Con aplicaciones en **Excel**

estadística aplicada a la administración y la economía

ALFREDO DÍAZ MATA

McGraw Hill

EBOOKS ACADÉMICOS

Tabla de áreas bajo la curva normal

z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0	0.004	0.008	0.012	0.016	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.091	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.148	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.17	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.195	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.219	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.258	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.291	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.334	0.3365	0.3389
1	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.377	0.379	0.381	0.383
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.398	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.437	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.475	0.4756	0.4761	0.4767
2	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.483	0.4834	0.4838	0.4842	0.4846	0.485	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.489
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.492	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.494	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.496	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.497	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.498	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.499	0.499
3.1	0.499	0.4991	0.4991	0.4991	0.4992	0.4992	0.4992	0.4992	0.4993	0.4993
3.2	0.4993	0.4993	0.4994	0.4994	0.4994	0.4994	0.4994	0.4995	0.4995	0.4995
3.3	0.4995	0.4995	0.4995	0.4996	0.4996	0.4996	0.4996	0.4996	0.4996	0.4997
3.4	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4998

Estadística aplicada a la administración y la economía

Estadística aplicada a la administración y la economía

Alfredo Díaz Mata
Universidad Nacional Autónoma de México
Facultad de Contaduría y Administración

Revisión técnica

Jorge Cardiel Hurtado
Universidad Nacional Autónoma de México
Facultad de Contaduría y Administración

Jesús Zepeda Orozco
Universidad Nacional Autónoma de México
Facultad de Contaduría y Administración

McGraw Hill

MÉXICO • BOGOTÁ • BUENOS AIRES • CARACAS • GUATEMALA • MADRID • NUEVA YORK
SAN JUAN • SANTIAGO • SAO PAULO • AUCKLAND • LONDRES • MILÁN • MONTREAL
NUEVA DELHI • SAN FRANCISCO • SINGAPUR • ST. LOUIS • SIDNEY • TORONTO

Director general México: Miguel Ángel Toledo Castellanos
Editor sponsor: Jesús Mares Chacón
Coordinadora editorial: Marcela Rocha Martínez
Editora de desarrollo: Karen Estrada Arriaga
Supervisor de producción: Zeferino García García

ESTADÍSTICA APLICADA A LA ADMINISTRACIÓN Y LA ECONOMÍA
Primera edición

Prohibida la reproducción total o parcial de esta obra,
por cualquier medio, sin la autorización escrita del editor.

Mc Graw Hill Educación

DERECHOS RESERVADOS © 2013, respecto de la primera edición por:
McGRAW-HILL/INTERAMERICANA EDITORES, S.A. DE C.V.
*A Subsidiary of **The McGraw-Hill** Companies, Inc.*
Prolongación Paseo de la Reforma 1015, Torre A,
Pisos 16 y 17, Colonia Desarrollo Santa Fe,
Delegación Álvaro Obregón,
C.P. 01376, México, D.F.
Miembro de la Cámara Nacional de la Industria Editorial Mexicana, Reg. Núm. 736

ISBN: 978-607-15-0846-1

All rights reserved

1234567890 1245678903

Impreso en México *Printed in Mexico*

The McGraw-Hill Companies

Dedicatoria

A la princesa Ale, al príncipe Bubú, y a la mamá de los pollitos, la Alma de mi alma.

Contenido

Acerca del autor xix
Prefacio xx
Agradecimientos xxii

capítulo 1

Introducción **1**
1.1 Aplicaciones de la estadística 1
1.2 Estadística descriptiva y estadística inferencial 3
1.3 Poblaciones y muestras; parámetros y estadísticos 3
1.4 Método estadístico 4
1.5 Datos estadísticos: variables y su clasificación 6
1.6 Escalas de medición 6
1.7 Variables continuas y variables discontinuas (o discretas) 8
1.8 Uso de computadoras en estadística 8
1.9 Excel 2007 9
1.10 Resumen 10

capítulo 2

Presentación de datos: tablas y gráficas **11**
2.1 Principales elementos de tablas y gráficas 11
2.2 Tablas 12
2.2.1 Series simples, series de datos y frecuencias, y series de clases y frecuencias 12
2.2.1.1 Series simples 14
2.2.1.2 Series de datos y frecuencias 16
2.2.1.3 Series de clases y frecuencias 17
2.2.1.4 Construcción de tablas de clases y frecuencias 18
2.2.2 Tablas de frecuencias para datos cualitativos 20
2.2.3 Frecuencias absolutas, relativas y acumuladas 20
2.2.4 Tablas de doble entrada o de clasificación cruzada o de contingencias 21
2.2.5 Uso de Excel. Construcción de distribuciones de frecuencias 22
2.3 Gráficas (con Excel) 26
2.3.1 Histogramas 26

2.3.2 Gráficas de líneas 29
2.3.3 Histogramas y polígonos de frecuencias 29
2.3.4 Gráficas circulares 31
2.3.5 Otras aplicaciones 32
2.4 Resumen 37
2.5 Ejercicios adicionales 37

capítulo 3

Medidas 44
3.1 Medidas de posición 44
3.1.1 Media aritmética 45
3.1.2 Media ponderada 47
3.1.3 Media armónica 48
3.1.4 Media geométrica 50
3.1.5 Medias o promedios móviles 51
3.1.6 Mediana 52
3.1.7 Moda 55
3.1.8 Percentiles 57
3.1.8.1 Cuartiles 58
3.1.8.2 Deciles 59
3.1.9 Relación entre la media, la mediana y la moda 60
3.2 Medidas de dispersión 67
3.2.1 Rango 68
3.2.2 Desviación media 69
3.2.3 Desviación intercuartílica 71
3.2.4 Varianza y desviación estándar 71
3.2.4.1 Uso de la varianza y la desviación estándar muestrales como estimadores 74
3.2.5 Aplicaciones comunes de la desviación estándar 74
3.2.5.1 Coeficiente de variación 74
3.2.5.2 Teorema de Chebyshev 75
3.2.5.3 Desviación estándar y distribución normal 76
3.3 Medidas de composición: la proporción 81
3.4 Medidas de forma: momentos 82
3.4.1 Tercer momento respecto a la media y el coeficiente de sesgo 83
3.4.2 Cuarto momento respecto a la media y el coeficiente de curtosis o apuntamiento 83
3.5 Funciones estadísticas de Excel y el complemento “Análisis de datos” 89
3.5.1 Estadística descriptiva 90
3.5.2 Media móvil 91
3.5.3 Percentil y jerarquía 92
3.6 Resumen 94
3.7 Fórmulas del capítulo 94
3.8 Ejercicios adicionales 96

capítulo 4

Introducción a la teoría de la probabilidad 108
4.1 Teoría de conjuntos y teoría de la probabilidad 109
4.2 Conceptos básicos, terminología y notación 112
4.2.1 Conceptos importantes 112

4.3 Técnicas de conteo, permutaciones y combinaciones 116
4.4 Interpretaciones de la probabilidad 120
4.4.1 Interpretación teórica o clásica 120
4.4.2 La probabilidad como frecuencia relativa 122
4.4.3 Interpretación subjetiva de la probabilidad 123
4.5 Axiomas de la probabilidad 124
4.5.1 Axioma sobre los posibles valores de la probabilidad 124
4.5.2 Axioma sobre la suma de las probabilidades de los eventos de un espacio muestral 124
4.5.3 Axioma sobre la probabilidad de ocurrencia de dos a más eventos mutuamente excluyentes 125
4.6 Regla de la suma de probabilidades 126
4.7 Probabilidad condicional 129
4.8 Independencia estadística 132
4.9 Regla de la multiplicación de probabilidades 133
4.9.1 La regla de la multiplicación para eventos independientes 134
4.10 Regla de Bayes 136
4.11 Resumen 139
4.12 Fórmulas del capítulo 140
4.13 Ejercicios adicionales 140

capítulo 5

Distribuciones discretas (discontinuas) de probabilidad 146
5.1 Introducción 146
5.2 Distribuciones de probabilidad de variables aleatorias discretas 147
5.3 Media y varianza de una distribución de probabilidades 149
5.4 Distribución binomial 151
5.4.1 Media y varianza de la distribución binomial 153
5.4.2 Distribución binomial, tablas de probabilidades binomiales y Excel 154
5.5 Tres formas de presentar una distribución de probabilidad 156
5.6 Distribución de Poisson 158
5.6.1 Distribución de Poisson, tablas de probabilidades Poisson y Excel 159
5.6.2 Distribución de Poisson como aproximación de la distribución binomial 161
5.6.3 Media y varianza de la distribución de Poisson 162
5.7 Distribución hipergeométrica 163
5.7.1 Media y desviación estándar de la distribución hipergeométrica 165
5.8 Distribución multinomial 168
5.8.1 Media y desviación estándar de la distribución multinomial 169
5.9 Resumen 170
5.10 Fórmulas del capítulo 170
5.11 Ejercicios adicionales 171

capítulo 6

Distribuciones continuas de probabilidad 177
6.1 Área como medida de probabilidad 177
6.2 Distribución normal de probabilidad 178

6.2.1 Características de la distribución normal 178
6.2.2 Distribución normal estándar 179
6.2.3 Tabla de áreas bajo la curva normal 179
6.2.3.1 Excel y áreas bajo la curva normal 183
6.2.3.2 Determinación de *z* a partir del área o la probabilidad 184
6.2.4 Determinación de probabilidades para cualquier distribución normal 188
6.2.4.1 Excel y probabilidades para cualquier distribución normal 189
6.2.4.2 Determinación de valores de la variable, o de *z*, a partir del área o la probabilidad 192
6.3 Ajuste cuando se utiliza la distribución normal para evaluar probabilidades de una variable discreta (ajuste por discontinuidad) 195
6.4 Aproximación de distribuciones de probabilidad de variables discontinuas con la distribución normal 197
6.4.1 Aproximación de la distribución binomial con la distribución normal 197
6.4.2 Aproximación de la distribución de Poisson con la distribución normal 200
6.5 Distribución exponencial de probabilidad 202
6.5.1 Relación entre la distribución exponencial y la distribución de Poisson 204
6.6 Otras distribuciones de probabilidad continuas 204
6.7 Advertencia 205
6.8 Resumen 205
6.9 Fórmulas del capítulo 205
6.10 Ejercicios adicionales 206

CAPÍTULO 7

Muestreo y distribuciones muestrales 209
7.1 Introducción al muestreo 209
7.1.1 Parámetros, estadísticos y estimadores 210
7.1.2 Estimación de parámetros y pruebas de hipótesis 210
7.1.3 Estimaciones por punto y estimaciones por intervalo 211
7.1.4 Muestreo aleatorio y muestreo de juicio 211
7.1.5 Muestreo aleatorio y Excel 211
7.1.5.1 Generación de números aleatorios 212
7.1.5.2 Muestra 213
7.1.6 Muestras únicas y muestras múltiples 214
7.1.7 Muestras relacionadas y muestras independientes 214
7.1.8 Tipos de muestreo aleatorio 214
7.1.9 Etapas de un estudio por muestreo 215
7.1.10 Distribuciones muestrales 215
7.2 Distribución muestral de la media 216
7.2.1 Desarrollo 217
7.2.2 Tres conclusiones importantes que se desprenden de la distribución muestral de la media: el teorema central del límite 219
7.2.3 Fórmula del error estándar de la media y factor de corrección por población finita 219
7.2.4 Consideraciones adicionales sobre la distribución muestral de la media 220
7.2.5 Aplicaciones del análisis de la distribución muestral de la media 222
7.3 Distribución muestral de la proporción 224
7.3.1 Desarrollo 224
7.3.2 Tres conclusiones importantes sobre la distribución muestral de la proporción 226
7.3.3 Fórmula del error estándar de la proporción y factor de corrección por población finita 227
7.3.4 Consideraciones adicionales sobre la distribución muestral de la proporción 228

7.4 Distribución muestral de la varianza 229
7.4.1 Distribuciones muestrales sin reemplazo y con reemplazo 231
7.4.2 Estimadores insesgados y estimadores sesgados 233
7.5 Resumen 234
7.6 Fórmulas del capítulo 235
7.7 Ejercicios adicionales 235

capítulo 8

Estimación de parámetros 240
8.1 Estimaciones por punto y estimaciones por intervalo 240
8.2 Error de muestreo y errores que no son de muestreo 241
8.3 Propiedades de los estimadores 241
8.4 Estimación de una media con muestras grandes 242
8.4.1 Los 3 elementos de una estimación por intervalo 244
8.4.2 Estimador y parámetro estimado 245
8.4.3 Cuándo sí se puede utilizar la distribución normal para hacer estimaciones de parámetros, y cuándo no 245
8.4.4 Determinación del tamaño de muestra necesario para estimar una media 245
8.4.4.1 Cuando no se incluye el factor de corrección por población finita 245
8.4.4.2 Cuando sí se incluye el factor de corrección por población finita 246
8.5 Comparación de la estimación de parámetros con muestras grandes y muestras pequeñas 248
8.5.1 Distribución *t* de Student, su tabla de áreas y Excel 249
8.6 Estimación de una media con muestras pequeñas 250
8.6.1 La población se distribuye de forma normal y se conoce la desviación estándar de la población: estadístico de prueba, *z* 250
8.6.2 La población se distribuye de forma normal pero no se conoce la desviación estándar de la población: estadístico de prueba, *t* de Student 250
8.6.3 La población no se distribuye de forma normal 251
8.7 Estimación de una proporción 252
8.7.1 Determinación del tamaño de muestra para estimar una proporción 253
8.7.1.1 Cuando no se incluye el factor de corrección por población finita 253
8.7.1.2 Incluyendo el factor de corrección por población finita 254
8.8 Otros intervalos de confianza 256
8.8.1 Intervalos de confianza para la diferencia entre 2 medias poblacionales 256
8.8.2 Intervalos de confianza para la diferencia entre 2 proporciones poblacionales 257
8.8.3 Intervalos de confianza para el total de una población a partir de una media 258
8.8.4 Intervalos de confianza para el total de una población a partir de una proporción 259
8.9 Resumen 261
8.10 Uso de Excel para construir intervalos 262
8.11 Fórmulas del capítulo 263
8.12 Ejercicios adicionales 263

capítulo 9

Pruebas de hipótesis 269
9.1 Introducción 269
9.2 Planteamiento de las hipótesis 270

9.3 Errores tipo I y tipo II 270
9.4 Procedimiento para realizar pruebas de hipótesis 273
9.5 Elaboración de una gráfica 274
9.6 Pruebas de 1 y de 2 extremos. Regiones de aceptación y de rechazo 275
9.6.1 Pruebas de 2 extremos o colas 275
9.6.2 Prueba de hipótesis de la cola inferior o del extremo izquierdo 275
9.6.3 Prueba de hipótesis de la cola superior o del extremo derecho 276
9.7 Métodos para realizar pruebas de hipótesis 278
9.7.1 Método del intervalo 278
9.7.2 Método del estadístico de prueba 278
9.7.3 Método del valor de la *P* 280
9.7.4 Resumen de los procedimientos para realizar pruebas de hipótesis con los 3 métodos 281
9.8 Prueba de hipótesis sobre una proporción poblacional 283
9.9 Resumen 286
9.10 Uso de Excel 286
9.11 Ejercicios adicionales 287

capítulo 10

Pruebas de hipótesis para 2 poblaciones 291
10.1 Panorama general de las pruebas de hipótesis 291
10.2 Pruebas de hipótesis sobre la diferencia entre 2 medias 293
10.2.1 Pruebas con muestras grandes e independientes 294
10.2.1.1 Cuando se conocen las varianzas de las 2 poblaciones 294
10.2.1.2 Cuando no se conocen las varianzas y no se asume que sean iguales 294
10.2.1.3 Cuando no se conocen las varianzas pero se asume que son iguales 295
10.2.2 Pruebas con muestras pequeñas e independientes, variables distribuidas normalmente 296
10.2.2.1 Cuando no se conocen las varianzas pero se asume que son iguales 296
10.2.2.2 Cuando no se conocen las varianzas y no se asume que sean iguales 297
10.2.3 Pruebas para muestras pareadas cuando no se conocen las varianzas pero no se necesita asumir que sean iguales 299
10.3 Pruebas de hipótesis sobre la diferencia entre 2 proporciones 303
10.4 Prueba para la diferencia entre 2 varianzas 305
10.4.1 Distribución *F* y Excel 307
10.5 Excel y pruebas de hipótesis para 2 muestras 308
10.6 Resumen 309
10.7 Fórmulas del capítulo 309
10.8 Ejercicios adicionales 310

capítulo 11

Pruebas de hipótesis con la distribución ji cuadrada 315
11.1 Introducción 315
11.2 Distribución ji cuadrada (χ^2) 316
11.3 Tablas de áreas bajo la curva de la distribución ji cuadrada 317
11.3.1 Excel y la tabla de áreas para χ^2 317
11.4 Pruebas de hipótesis para la varianza de una población 318

11.5 Distribución ji cuadrada a partir de frecuencias observadas y frecuencias esperadas 319
11.6 Pruebas para una proporción con z y con χ^2 321
11.6.1 Prueba de una proporción con z 321
11.6.2 Prueba de una proporción con χ^2 322
11.7 Prueba para la diferencia entre 2 proporciones con z y con χ^2 322
11.7.1 Prueba para la diferencia entre 2 proporciones con z 323
11.7.2 Prueba para la diferencia entre 2 proporciones con χ^2 323
11.8 Relación entre las pruebas de hipótesis para proporciones con z y con χ^2 324
11.9 Prueba para la diferencia entre n proporciones 325
11.10 Pruebas de bondad de ajuste a distribuciones teóricas 326
11.10.1 Pruebas de bondad de ajuste a una distribución normal 327
11.10.2 Pruebas de bondad de ajuste a una distribución Poisson 329
11.10.3 Pruebas de bondad de ajuste a una distribución binomial 330
11.11 Pruebas de bondad de ajuste entre distribuciones empíricas 334
11.12 Pruebas sobre la independencia entre 2 variables 336
11.13 Pruebas paramétricas y pruebas no paramétricas 337
11.14 Excel y la distribución ji cuadrada 338
11.14.1 Función Distr.Chi 338
11.14.2 La función Prueba.Chi 338
11.15 Resumen 339
11.16 Fórmulas del capítulo 339
11.17 Ejercicios adicionales 340

capítulo 12

Análisis de varianza 345
12.1 Introducción 345
12.2 Suposiciones en que se basan las técnicas de análisis de varianza 348
12.3 El diseño completamente aleatorizado de un factor 349
12.4 Procedimiento para el ANOVA con el diseño completamente aleatorizado de un factor 352
12.5 Excel y ANOVA de un factor 354
12.6 Comparaciones múltiples entre pares de medias de tratamiento 356
12.7 Análisis de varianza de dos factores 357
12.8 Excel y ANOVA de dos factores 360
12.9 Análisis de varianza de dos factores con interacción 362
12.10 Excel y ANOVA de dos factores con interacción 366
12.11 Resumen 369
12.12 Fórmulas del capítulo 369
12.13 Ejercicios adicionales 370

capítulo 13

Análisis de regresión y correlación lineal simple 374
13.1 Ecuación y recta de regresión 375
13.2 Método de mínimos cuadrados 377

13.2.1 Derivación algebraica de las ecuaciones normales 378
13.2.2 Derivación de las ecuaciones normales mediante derivadas parciales 379
13.3 Determinación de la ecuación de regresión 380
13.3.1 Despeje simultáneo de *a* y *b* en las 2 ecuaciones normales 380
13.3.2 Resolución simultánea de las 2 ecuaciones normales 381
13.3.3 Resolución mediante sumas de cuadrados 382
13.3.4 Uso de Excel 383
13.4 Modelo de regresión y sus supuestos 386
13.5 Sumas de cuadrados en el análisis de regresión 387
13.6 Desviación estándar de regresión 388
13.7 Inferencias estadísticas sobre la pendiente β_1 389
13.7.1 Pruebas de hipótesis sobre la pendiente β_1 389
13.7.1.1 Prueba sobre la pendiente β_1 utilizando la *t* de Student 390
13.7.1.2 Prueba sobre la pendiente utilizando la *F* de Fisher 390
13.7.2 Estimación por intervalo de β_1 391
13.8 Uso de la ecuación de regresión para estimación y predicción 391
13.8.1 Estimación por intervalo de *y* para valores dados de *x* 391
13.8.2 Pronósticos de *y* para valores dados de *x* 392
13.9 Recapitulación del análisis de regresión lineal simple 393
13.10 Análisis de correlación 399
13.10.1 Coeficiente de correlación y Excel 400
13.10.2 Momento-producto de Pearson, otra manera de interpretar el coeficiente de correlación 402
13.10.3 Prueba de hipótesis sobre el coeficiente de correlación 403
13.10.4 Correlación serial o autocorrelación 403
13.10.4.1 Prueba de hipótesis sobre el coeficiente de correlación serial 405
13.11 Resumen 406
13.12 Fórmulas del capítulo 407
13.13 Ejercicios adicionales 408

capítulo 14

Análisis de regresión lineal múltiple 411
14.1 Modelo de regresión lineal múltiple y sus supuestos 411
14.2 Obtención de la ecuación de regresión lineal múltiple 412
14.3 Multicolinealidad y las variables que mejor se relacionan con la variable dependiente: uso de la matriz de correlaciones 415
14.4 Evaluación de la ecuación de regresión 418
14.4.1 Evaluación de la ecuación de regresión mediante el coeficiente de determinación múltiple 418
14.4.2 Evaluación de la ecuación de regresión mediante el análisis de varianza y la prueba *F* 419
14.4.3 Inferencias sobre coeficientes de regresión parciales individuales 422
14.4.4 Análisis de residuales 425
14.5 Uso del modelo de regresión lineal múltiple 431
14.5.1 Intervalos de confianza para los pronósticos 432
14.5.2 Intervalos de confianza para estimaciones de la media de una subpoblación de valores *y* 432
14.6 Variables independientes cualitativas 433
14.7 Regresión por pasos 437

14.7.1 Eliminación posterior 437
14.7.2 Regresión por pasos mediante selección previa 438
14.8 Resumen 441
14.9 Fórmulas del capítulo 441
14.10 Ejercicios adicionales 442

capítulo 15

Números índice 450
15.1 Usos de los números índice 450
15.2 Números índice simples 451
15.3 Números índice agregados 453
15.4 Números índice agregados de Laspeyres, de Paasche e ideal de Fischer 454
15.4.1 Índice de Laspeyres 454
15.4.2 Índice de Paasche 455
15.4.3 Índice ideal de Fischer 456
15.5 Números índices en cadena 459
15.5.1 Números índice en cadena y rendimientos bursátiles 460
15.6 Índices para propósitos especiales 463
15.6.1 Índices de precios al consumidor y al productor 463
15.6.1.1 El Índice Nacional de Precios al Consumidor 463
15.6.1.2 Cálculo de la inflación mediante el INPC 465
15.6.1.3 Cambio de periodo base 465
15.6.1.4 Deflación de series de tiempo con el INPC 466
15.6.1.5 El Índice Nacional de Precios al Productor 468
15.6.2 Índices bursátiles 468
15.7 Números índices y Excel 473
15.8 Resumen 473
15.9 Fórmulas del capítulo 473
15.10 Ejercicios adicionales 474

capítulo 16

Análisis de series de tiempo 476
16.1 Modelo clásico de series de tiempo 476
16.2 Análisis gráfico de la tendencia 478
16.3 Tendencia secular 479
16.3.1 Suavización con promedios móviles exponenciales 479
16.3.1.1 Pronósticos con promedios móviles exponenciales 480
16.3.2 Ajuste de una recta con mínimos cuadrados 481
16.3.3 Ajuste de una función exponencial con mínimos cuadrados 482
16.3.4 Ajuste de una parábola con mínimos cuadrados 486
16.4 Variaciones estacionales 492
16.4.1 Cálculo de índices estacionales 492
16.4.2 Desestacionalización de series de tiempo 498
16.4.3 Pronósticos con índices estacionales 499
16.5 Variaciones cíclicas 500
16.6 Resumen 506

16.7 Fórmulas del capítulo 506
16.8 Ejercicios adicionales 506

capítulo 17

Pruebas estadísticas no paramétricas 511
17.1 Pruebas de hipótesis no paramétricas en este libro 512
17.2 Prueba de rachas para aleatoriedad de Wald-Wolfowitz 513
17.2.1 Características 514
17.2.2 Excel y el cálculo de probabilidades para la prueba de rachas de Wald-Wolfowitz 516
17.3 Prueba de los signos 518
17.3.1 Características 518
17.3.2 Prueba del signo para una muestra pequeña 519
17.3.3 Prueba del signo para una muestra grande (aproximación normal) 521
17.3.4 Prueba del signo para 2 muestras apareadas pequeñas 521
17.3.5 Prueba del signo para 2 muestras apareadas grandes (aproximación normal) 522
17.4 Prueba de rangos con signo de Wilcoxon 526
17.4.1 Características 527
17.4.2 Prueba de rangos con signo de Wilcoxon para una muestra pequeña 527
17.4.3 Excel y la prueba de rangos con signo de Wilcoxon 528
17.4.4 Prueba de rangos con signo de Wilcoxon para una muestra grande (aproximación normal) 529
17.4.5 Prueba de rangos con signo de Wilcoxon para 2 muestras apareadas pequeñas 530
17.4.6 Prueba de rangos con signo de Wilcoxon para 2 muestras apareadas grandes (aproximación normal) 531
17.5 Prueba *U* de Mann-Whitney para 2 muestras independientes 532
17.5.1 Características 532
17.6 Prueba de suma de rangos de Kruskal-Wallis para más de 2 medias 536
17.7 Prueba de Friedman para diseños en bloques aleatorizados 539
17.8 Coeficiente de correlación por rangos de Spearman 541
17.9 Resumen 543
17.10 Fórmulas del capítulo 543
17.11 Ejercicios adicionales 545

Apéndices 552
Apéndice 1 Tabla de probabilidades nominales 552
Apéndice 2 Tabla de probabilidades de Poisson 560
Apéndice 3 Tabla de áreas bajo la distribución *t* de Student 566
Apéndice 4 Tablas de la distribución *F* 567
Apéndice 5 Tabla de áreas bajo la distribución ji cuadrada 569
Apéndice 6 Valores críticos de la *T* de Wilcoxon 570
Apéndice 7 Tabla para la prueba de Mann Whitney y de Daniel y Terrell 571

Glosario 574
Respuestas a los ejercicios nones 578
Índice analítico 604

Acerca del autor

El doctor Alfredo Díaz Mata ha sido profesor de matemáticas en la Facultad de Contaduría y Administración de la Universidad Nacional Autónoma de México durante los últimos 35 años. Tiene estudios de licenciatura en administración con especialización en estadística aplicada, así como maestría y doctorado en ciencias de la administración, todos ellos por la UNAM.

Ha traducido más de 30 textos del inglés al español sobre temas de matemáticas, administración y finanzas. Es autor o coautor de una decena de libros, entre los que destaca *Matemáticas financieras*, cuya primera edición se publicó hace unos 25 años por McGraw-Hill y el cual va en su quinta edición.

Su experiencia laboral incluye análisis de estudios de mercado en Procter & Gamble de México y administración de sueldos y salarios en American Express México.

Prefacio

Se ha extendido el uso de métodos estadísticos para el análisis y resolución de numerosos problemas prácticos que se presentan en muchas disciplinas, en particular, en las áreas de la administración, la economía y las ciencias sociales en general. Esta diversificación ha sido muy extendida, sobre todo, a partir de la amplia disponibilidad de computadoras personales y de software para análisis estadístico, que no es exagerado decir que se ha vuelto universal.

Este texto pretende acercar a los estudiantes universitarios de dichas áreas a este conjunto de técnicas estadísticas y busca hacerlo de manera que se comprenda tanto la teoría y la lógica de los métodos como su aplicación en la resolución de problemas prácticos. Por ello, se utilizan explicaciones y ejemplos detallados que, sin perder el rigor que la disciplina exige, faciliten al estudiante la comprensión de la teoría y la técnica, mientras se mantiene la conexión que ambas tienen en la práctica.

Dada su difundida utilización en los ámbitos académico y profesional, el libro hace también hincapié en el uso de Excel de Microsoft como un auxiliar muy útil en la aplicación de las técnicas estadísticas, ya que cuenta con numerosas características que facilitan el trabajo en buena medida. Es por ello que se presentan abundantes ejemplos y explicaciones sobre la utilización de esta herramienta.

El libro se divide en 17 capítulos. En el primero se explica qué es la estadística y se plantea un panorama general de sus principales técnicas. Le sigue el capítulo 2 que trata sobre la elaboración de tablas y gráficas y el capítulo 3 que aborda las principales medidas estadísticas: de posición, de dispersión, de composición y de forma.

En el capítulo 4 se presenta una introducción a la teoría de la probabilidad, seguido de dos capítulos sobre distribuciones de probabilidad: las discretas se estudian en el capítulo 5 y las continuas, en el capítulo 6. De éstas destaca, por supuesto, la muy útil distribución normal que tiene forma de campana.

El capítulo 7, que se apoya en los capítulos 4, 5 y 6, se ocupa de las distribuciones muestrales y resume la mayor parte de la teoría fundamental en la que se basan casi todos los capítulos restantes y que es la que se requiere para aplicar las dos principales técnicas de la inferencia estadística: las estimaciones de parámetros, que se abordan en el capítulo 8, y las pruebas de hipótesis, que se estudian en el capítulo 9.

En los capítulos 9 y 10 se desarrollan los conceptos principales sobre pruebas de hipótesis para una población y para dos poblaciones y, al igual que en el resto del libro, se resuelven numerosos ejemplos.

Las pruebas de hipótesis forman el grueso de la parte restante del libro y, en la sección 10.1, se presenta un panorama general de las pruebas de hipótesis que se revisan en los capítulos restantes. Se sugiere revisar con especial atención este panorama general para tener una idea global de este amplio campo de análisis estadístico.

En el capítulo 11 se ilustran los procedimientos para realizar pruebas de hipótesis de bondad de ajuste, de independencia y de homogeneidad, utilizando la distribución ji cuadrada, en tanto que el capítulo 12 se ocupa del tema de análisis de varianza, cuya utilidad primordial radica en la realización de pruebas de hipótesis para más de dos poblaciones.

En los capítulos 13 y 14 se revisan los importantes y ampliamente utilizados temas de análisis de regresión y correlación lineal, tanto simple como múltiple, en tanto que en el capítulo 15 se revisan los temas de números índices, en el capítulo 16 se examina el análisis de series de tiempo y, para terminar, el capítulo 17 aborda otro tema de especial importancia para las ciencias sociales en general, las pruebas de hipótesis no paramétricas.

Cada capítulo comienza con una introducción que presenta una perspectiva de su contenido para, después, tratar los diferentes temas básicos correspondientes con base en explicaciones sencillas pero rigurosas de conceptos y la resolución de abundantes ejemplos, tanto "a mano" (en los cuales se asume el uso de calculadoras electrónicas), como utilizando el Excel de Microsoft.

El libro cuenta con una gran cantidad de ejercicios. Dado que cada capítulo está dividido en secciones, se han insertado varias series de ejercicios por sección y, además, al final de cada capítulo se incluye otra sección de ejercicios la cual cubre todos los temas y se denomina "ejercicios adicionales". Los ejercicios por sección pueden servir a diversos propósitos. En primer lugar, pueden ser resueltos por los estudiantes para asegurarse de que han comprendido las técnicas estadísticas correspondientes y su aplicación; es por ello que, en un apéndice al final del libro, se presentan las respuestas de todos los ejercicios por sección con numeración non, lo cual permite verificar que el resultado que se obtiene es el correcto. Sin embargo, aunque se muestran las respuestas de gran parte de los ejercicios nones, no se incorporan aquellas que implican mucho espacio y cuya resolución es muy sencilla; tampoco se ofrecen las respuestas de los ejercicios que pueden tener varias respuestas, como los que se basan en el trazado manual de rectas o los que deben resolverse con datos que se deben obtener de internet al momento de resolverlos.

Por otra parte, los ejercicios por sección pueden servir para que los profesores elaboren exámenes y para que realicen prácticas adicionales en el salón de clases. Los ejercicios adicionales que se incluyen al final de cada capítulo también se pueden utilizar para los mismos fines.

Agradecimientos

La elaboración de este libro no hubiera sido posible sin la ayuda de numerosas personas, tanto estudiantes como profesores, cuya lista completa me resulta ya imposible reunir. Con una disculpa para quienes no quedan incluidos aquí, quiero mencionar a los profesores Jorge Cardiel Hurtado, Mario Gabriel Gómez Urquiza y Álvarez de la Cadena y Jesús Zepeda Orozco.

Por otra parte, también deseo agradecer el apoyo que me prestaron en esta labor los estudiantes Alejandra Selene Castillo Valdez, Ana Paola Castillo González, Ana María Rincón Botero, Irving Bustamante Herrera, Josué Chávez y Juan Andrés Pérez Celis.

Finalmente, quiero gratificar también la ayuda, sin la cual tampoco hubiera visto la luz este libro, de los editores de McGraw-Hill/Interamericana: Jesús Mares Chacón, Marcela Rocha Martínez y Karen Estrada Arriaga.

capítulo 1

Introducción

Sumario

1.1 Aplicaciones de la estadística
1.2 Estadística descriptiva y estadística inferencial
1.3 Poblaciones y muestras; parámetros y estadísticos
1.4 Método estadístico
1.5 Datos estadísticos: variables y su clasificación
1.6 Escalas de medición
1.7 Variables continuas y variables discontinuas (o discretas)
1.8 Uso de computadoras en estadística
1.9 Excel 2007
1.10 Resumen

En el lenguaje cotidiano es común hablar de *estadística* en 2 sentidos: por un lado se habla de *estadística* para referirse a un conjunto determinado de datos, por ejemplo, *estadísticas de población* cuando se interpretan los resultados de un censo o *estadísticas de ventas* relacionadas con datos históricos sobre ventas.

El otro sentido de la palabra se refiere a una disciplina matemática con la posibilidad de abordarse desde varias perspectivas. La *estadística* es, en este segundo sentido, un área de las matemáticas —objeto formal de estudio en diversos niveles académicos desde secundaria hasta posgrado— sobre la cual los especialistas y estudiosos siguen trabajando y produciendo avances, existen textos con títulos sugerentes como *Estadística matemática*. De manera igualmente formal, la estadística también se estudia desde el punto de vista práctico, de su aplicación. Y es tal su utilidad en la administración de empresas públicas y privadas como en todo tipo de organizaciones, que desde hace muchos años existen cursos de estadística aplicada en niveles medio superior, superior y de posgrado. El título de este libro, *Estadística aplicada a administración y economía*, refleja esta orientación y enfatiza el hecho de que su objetivo no es la estadística matemática sino la aplicada.

Una manera de describir la **estadística**, en este sentido de disciplina aplicada, consiste en considerarla como un conjunto de técnicas para el análisis de datos, éstos suelen ser números aunque, como se verá después, existen datos estadísticos que no son precisamente números sino que son categorías. En la sección siguiente se presenta un panorama general de las principales técnicas de estadística para mejorar la idea de cuál es el campo de estudio de la estadística en este sentido. En las secciones 1.2 a 1.8 se revisarán diversos conceptos cuya comprensión es importante para tener una mejor idea, desde el inicio, del tipo de análisis que se realiza y de las posibles aplicaciones de la estadística. En las secciones finales se presentarán un resumen del capítulo y un conjunto de ejercicios.

Estadística. Disciplina matemática considerada como un conjunto de técnicas para el análisis de datos.

1.1 Aplicaciones de la estadística

Antes se mencionó que la estadística es un conjunto de técnicas para el análisis de datos. Esta sencilla definición sirve para puntualizar los 2 componentes principales de esta disciplina: análisis y datos. Los **datos** son la materia prima de la estadística, en tanto que las **técnicas de análisis** son los mecanismos mediante los cuales dichos datos se convierten en información útil. Debido a que los datos son un ingrediente tan importante en el análisis estadístico, en las secciones 1.5 y 1.6 se revisarán con detalle varios aspectos relacionados con ellos: los datos como mediciones, el concepto de variable y otros.

Datos. Materia prima de la estadística.
Técnicas de análisis. Mecanismos mediante los cuales se convierten los datos en información útil.

En los siguientes párrafos se revisan brevemente las técnicas de análisis estadístico que constituyen la estructura de este libro, y que dan idea de las aplicaciones prácticas de esta disciplina.

Una de las aplicaciones más comunes de la estadística es la elaboración de tablas y gráficas que sirven para recopilar, organizar y presentar datos. La elaboración de tablas y gráficas incluye la revisión del tipo

de datos que se analizan, las fuentes donde se obtienen y las formas en que pueden resumirse para facilitar su interpretación.

La principal ventaja de las tablas es que permiten presentar conjuntos grandes de datos en forma compacta. En cuanto a las gráficas, es bien conocida su virtud de ofrecer información fácilmente asimilable. Sin embargo, es evidente que estas herramientas ofrecen sus ventajas sólo cuando se les elabora en forma adecuada. Es precisamente aquí en donde intervienen las técnicas estadísticas para la elaboración de tablas y gráficas que se analizarán en el capítulo 2.

Otra de las aplicaciones importantes y frecuentes de la estadística es la utilización de números simples para resumir características importantes de conjuntos de datos. Estos números simples que representan características propias de conjuntos de datos son denominados **medidas** e indudablemente la más conocida de ellas es el promedio aritmético, también llamado *media aritmética* o simplemente *media*. Ejemplos de media son: el promedio de kilómetros que un automóvil recorre por cada litro de gasolina que consume; el promedio de edad de un grupo de individuos, media de la que casi toda persona tiene una idea bastante buena, aunque sólo sea intuitiva. Para comprobar esto bastaría preguntar a alguien cuál es el promedio de edad de un grupo determinado de personas a fin de que pueda fácilmente aproximar una respuesta, aun cuando no sepa definir con precisión el procedimiento para hacer los cálculos. Otra medida que se utiliza con frecuencia es la "proporción", entrecomillamos el término porque tiene un sentido técnico del cual se hablará en la sección correspondiente a medidas. ¿Quién no ha hablado o escuchado acerca de, por ejemplo, la proporción de hombres o mujeres en un determinado conjunto de personas? Estas 2 medidas, la media y la proporción, junto con otras más que se revisarán en el capítulo 3 se utilizan para analizar conjuntos de datos en forma sintetizada.

Medidas. Números simples que representan características de conjuntos de datos.

Un área muy importante del análisis estadístico es el **muestreo**: se escoge una cantidad relativamente reducida de elementos representativos de una población numerosa para inferir conclusiones acerca de la totalidad de la población; a esta parte de la estadística se le conoce como **estadística inferencial**. Los usos y aplicaciones del muestreo son numerosos y variados, dada la enorme complejidad de las organizaciones modernas, es una de las áreas que más se estudian y que más se utilizan. Por esto no es de extrañar que una parte considerable del libro se dedique a ella. Asimismo se examina la teoría de la probabilidad porque es el fundamento del muestreo; en estudios por muestreo es común que se hagan afirmaciones como:

Muestreo. Cantidad relativamente reducida de elementos representativos de una población.
Estadística inferencial. Parte de la estadística que por medio del muestreo infiere conclusiones acerca de la totalidad de una población.

> …con base en los resultados obtenidos con la muestra, se puede afirmar que el promedio de ingresos mensuales de los hogares de la población de donde se obtuvo la muestra es de $8 760, y se tiene una probabilidad de 95% de estar en lo correcto al afirmarlo.

Es claro que como no se estudia la totalidad de la población, no se puede tener la certeza de la veracidad de lo que se dice; sin embargo, por medio de la estadística se pueden asociar criterios de probabilidad a las afirmaciones, y esas probabilidades permiten evaluar qué tan útiles son los resultados obtenidos.

Una de las técnicas más importantes del muestreo es la conocida como **pruebas de hipótesis**. Estas pruebas son procedimientos a través de los cuales se intenta verificar si ciertas suposiciones acerca de la población son ciertas o no. Por ejemplo, se pueden tener ciertas bases para suponer que la proporción de personas de sexo masculino es de 25% en determinada población. También puede suceder que se desee verificar esa suposición a través de un estudio muestral. Una vez realizado, y con base en los resultados obtenidos, se puede tomar la decisión de aceptar o rechazar la hipótesis (suposición) planteada.

Pruebas de hipótesis. Procedimientos a través de los cuales se trata de verificar si ciertas suposiciones acerca de la población son ciertas o no.

Otra aplicación de la estadística también se ocupa de la relación entre 2 o más conjuntos de datos. Por ejemplo, se revisa el posible grado de asociación entre los ingresos familiares y el consumo de determinado producto; esta relación, si puede establecerse, sería de utilidad para las empresas que venden el producto. Otra aplicación importante de esta técnica es el **análisis de regresión y correlación**, utilizado en economía. Aquí se estudia la relación que pudiera existir entre la inversión y otras variables como el producto interno bruto, las tasas de interés, el dinero circulante, etc. Se estudia también el **análisis de series de tiempo**, es decir, el comportamiento que tienen ciertos indicadores con el transcurso de un periodo. Estos indicadores pueden ser las ventas de un negocio, la tasa de inflación o los salarios de algún tipo de trabajadores.

Análisis de regresión y correlación. En economía, estudia la relación que pudiera existir entre la inversión y otras variables.
Análisis de series de tiempo. Comportamiento que tienen ciertos indicadores en el transcurso de un periodo.

Con los **números índice** se estudian las variaciones que sufren determinadas mediciones (comúnmente precios y cantidades de artículos) de un periodo a otro, o durante diversos periodos. Uno de los números índice más conocidos es el Índice Nacional de Precios al Consumidor que mide las variaciones de los precios de un conjunto de artículos (la canasta básica) de un día a otro, de un mes a otro o de un año a otro. Otro número índice importante y famoso es el *índice de la bolsa*, que en México se conoce como Índice de Precios y Cotizaciones (IPC) de la Bolsa Mexicana de Valores y mide cuánto subió o bajó una muestra representativa de todas las acciones que se negocian en la Bolsa. Este IPC aparece con frecuencia en los medios de comunicación para sustentar afirmaciones como "bajó la bolsa" o "la bolsa subió 10% en el último mes".

Números índice. Se utilizan para estudiar las variaciones que sufren determinadas mediciones de un periodo a otro, o durante diversos periodos.

Una parte de la estadística cuya importancia tiene una tendencia creciente en los últimos tiempos es el **análisis multivariado**. En esta área se revisan diversas técnicas empleadas en el análisis de más de 2 variables o mediciones. En este sentido se mencionó al análisis de regresión que estudia, por ejemplo, la relación entre los ingresos y el consumo. Esta relación es entre 2 variables pero en muchos estudios prácticos surge la necesidad de considerar más de 2 variables, y es aquí en donde interviene el análisis multivariado.

Análisis multivariado. Parte de la estadística que revisa varias técnicas empleadas en el análisis de más de 2 variables o mediciones.

Por supuesto, dentro de cada una de las áreas mencionadas en los párrafos anteriores se abarcan diversos procedimientos de análisis que se aplican en la forma esbozada a circunstancias específicas.

Después de revisada esta sección, se sugiere al estudiante leer y analizar el contenido de este libro y se recomienda revisar la tabla de contenido de cualquier otro texto similar para repasar de manera más amplia este conjunto de técnicas estadísticas y también para comprobar las grandes coincidencias que existen en la literatura sobre el área.

1.2 Estadística descriptiva y estadística inferencial

Esta división de la estadística en 2 grandes áreas de estudio es útil para comprender mejor el papel de esta disciplina. La **estadística descriptiva** se ocupa del análisis de los datos con el propósito de recopilarlos, organizarlos, resumirlos, etc., con las técnicas mencionadas antes pero sin incluir el uso de muestras para hacer inferencias. Es decir, en la estadística descriptiva se aplican las técnicas de recopilación de datos (que pueden corresponder a muestras pero sin intención de utilizarlas para hacer suposiciones acerca de la población de donde se obtienen). Se utilizan también para la elaboración de tablas y gráficas; el cálculo de medidas simples (números únicos) para resumir conjuntos grandes de datos (media aritmética, proporción, varianza y otras); el análisis de regresión y correlación (sin embargo, en el caso particular del análisis de regresión y correlación, también se pueden hacer estudios mediante muestras con el propósito de hacer inferencias y que, por ello, corresponderían a la estadística inferencial); el de series de tiempo, y el de números índices.

Estadística descriptiva. Se ocupa del análisis de los datos sin utilizar muestras para hacer inferencias.

Por su parte, la **estadística inferencial** es la que se ocupa del análisis de muestras con el objeto de obtener conclusiones (inferencias) acerca de la población de donde se obtienen los datos.

Estadística inferencial. Se ocupa del análisis de muestras para obtener conclusiones (inferencias) acerca de la población de donde se obtienen los datos.

Como puede leerse, la diferencia entre las 2 áreas de la estadística no es enteramente clara pero es útil para comprender mejor la materia. Resulta también conveniente tener presente que, en sus orígenes, la estadística era del tipo descriptivo y que en tiempos recientes la estadística inferencial aumenta en importancia; actualmente ha llegado a constituir parte considerable de las técnicas que se aplican. La estadística inferencial es un área en la que se realizan constantemente estudios teóricos y prácticos para mejorar las técnicas existentes y para crear nuevas. Además, esta diferencia entre estadística descriptiva e inferencial conduce a la necesidad de distinguir claramente entre poblaciones y muestras, que es el tema de la siguiente sección.

1.3 Poblaciones y muestras; parámetros y estadísticos

Una **población** es el conjunto de todos los elementos o unidades de interés para un estudio determinado. Algunos ejemplos:

- El conjunto de las amas de casa que habitan en determinada ciudad sería una población si se tratara de determinar su nivel de aceptación para cierto producto de limpieza.

Población. Conjunto de todos los elementos o unidades de interés para un estudio determinado.

- El conjunto de las piezas fabricadas por cierta máquina sería una población si se buscara determinar el porcentaje de artículos defectuosos que produce.
- El total de las cuentas por cobrar de una empresa sería una población si se intentara determinar el porcentaje de cuentas morosas de esa empresa.
- El total de los habitantes de un país es la población que se considera cuando se lleva a cabo un censo de ese país, como el que se aplica en México cada década (el más reciente en 2010).
- El número total de hectáreas cultivadas de maíz en una región sería la población cuando se trata de determinar el rendimiento promedio de maíz por hectárea en esa región.

Como puede verse, el concepto de *población* es relativo pues su definición depende de lo que se pretende estudiar. Debe definirse con cuidado cuál es la población (o universo) que se desea estudiar. Otro aspecto muy importante a tener en cuenta para realizar estudios por muestreo es definir con claridad y precisión cuáles son los elementos de la población, de manera que no haya confusiones para determinar si algún caso específico pertenece o no a la población.

Muestra. Subconjunto de los elementos de una población.

Por otra parte, una **muestra** es un subconjunto de los elementos de una población. La principal característica que debe tener una muestra estadística útil es: ser representativa de la población de donde se extrae, porque el principal propósito de la obtención de muestras consiste en hacer inferencias sobre la población correspondiente. El método más común para obtener una muestra representativa consiste en elegir sus elementos al azar y se les denomina **muestras aleatorias**; siguiendo los ejemplos presentados, las muestras aleatorias correspondientes a las poblaciones descritas serían subconjuntos de amas de casa, subconjuntos de piezas fabricadas, subconjuntos de cuentas por cobrar, subconjuntos de habitantes y subconjuntos de hectáreas cultivadas de maíz.

Muestras aleatorias. Muestra representativa cuyos elementos son elegidos al azar.

Una vez identificada la muestra o la población con la que se trabajará deben recopilarse los datos que se usarán en el análisis. Esta recopilación de datos consiste, desde el punto de vista estadístico, en tomar una medición de cada uno de los elementos y después proceder al análisis con este conjunto de mediciones. En otras palabras, es importante tener presente la diferencia entre los elementos de la población o muestra y las mediciones que se hacen a cada uno de esos elementos, que son las que constituyen, precisamente, los datos materia prima de la estadística.

Así, en un estudio por muestreo, la población podría ser el conjunto de los hogares de cierta ciudad de donde se tomaría una muestra; después se indagaría el ingreso mensual de cada uno de los hogares de la muestra para calcular el promedio del ingreso mensual por hogar. Cada hogar sería un elemento de la muestra y su correspondiente ingreso sería la medición (o como se verá más adelante, el valor de la variable de interés). Finalmente, con ese promedio muestral se podría estimar (inferir) el promedio del ingreso mensual de todos los hogares de la población (la ciudad completa).

El ejemplo anterior sirve para ilustrar otros 2 conceptos muy importantes:

Estadístico muestral o estadístico. Medida de una muestra.
Parámetro. Medida de una población.

- **Estadístico muestral** o **estadístico.** Es una medida de una muestra, por ejemplo, el promedio del ingreso por hogar en la muestra.
- **Parámetro.** Es una medida de una población, por ejemplo, el promedio del ingreso por hogar en la población.

Esta diferenciación es muy importante porque no es lo mismo la media de la muestra que la media de la población porque, por lo general, en estadística inferencial se utiliza el estadístico (la media muestral) para inferir (estimar) el valor del parámetro (la media de la población) que en muchos casos se desconoce.

Así como se diferencian medias muestrales y medias poblacionales, se aplica la misma distinción a todas las demás medidas que se analizarán; en estudios por muestreo se puede hablar de una proporción muestral (el estadístico) y de una proporción poblacional (el parámetro), de una varianza muestral (el estadístico) y de una varianza poblacional (el parámetro), etcétera.

1.4 Método estadístico

La aplicación de las técnicas estadísticas para el análisis de datos puede resumirse en un método que consta de 5 etapas:

1. Recopilación de datos.
2. Organización de datos.
3. Presentación de datos.
4. Análisis de datos.
5. Conclusiones.

Se resumen en seguida los principales aspectos a considerar en cada una de estas etapas:

1. Recopilación de datos. Es importante tener presente la fuente de los datos con los que se trabaja. Es posible tener **datos de fuentes internas**, que son aquellos generados al interior de la organización, como pueden ser datos contables, registros de ventas, de personal, etc. Existen también los **datos de fuentes externas** que pueden obtenerse de otras personas u organizaciones. Los datos se clasifican, por un lado, si son **datos de fuentes primarias**, en cuyo caso son generados por quien los utiliza, como serían los datos de fuentes internas. Sin embargo, los datos de fuentes primarias también pueden conseguirse en el exterior por medio de encuestas (como las realizadas en investigación de mercados) o directamente a través de otras organizaciones que sean la fuente primaria. Por otro lado, están los **datos de fuentes secundarias** que se recogen de fuentes que no son los recopiladores originales de los datos. Un ejemplo de esta clase de datos son los que aparecen en compendios de datos económicos y financieros que reúnen en un solo volumen información publicada por diferentes organizaciones gubernamentales y privadas.

   Datos de fuentes internas. Datos que se generan al interior de la organización.
   Datos de fuentes externas. Datos que se pueden obtener de otras personas u organizaciones.
   Datos de fuentes primarias. Datos que son generados por quien los utiliza, como sería el caso de los datos de fuentes internas.
   Datos de fuentes secundarias. Datos que se obtienen de fuentes que no son los recopiladores originales de la información.

   Vale la pena mencionar algunas de las principales fuentes de datos publicados que son de uso común en el área de negocios y economía. A estas bases de datos se puede acceder a través de internet:

   - Banco de México, S.A. (Banxico): www.banxico.gob.mx
   - Nacional Financiera, S.A. (Nafinsa): www.nafin.com
   - Instituto Nacional de Estadística, Geografía e Informática (INEGI): www.inegi.gob.mx
   - Confederación de Cámaras Nacionales de Comercio, Servicios y Turismo (Concanaco): www.concanaco.com.mx
   - Cámara Nacional de la Industria de la Transformación (Canacintra): www.canacintra.org.mx
   - Publicaciones periódicas especializadas como las siguientes:

     a) Periódico *El Financiero*: www.elfinanciero.com.mx
     b) Periódico *El Economista*: www.eleconomista.com.mx
     c) Revista *Expansión*: www.cnnexpansion.com
     d) Las revistas publicadas por el Instituto Mexicano de Ejecutivos de Finanzas (IMEF): www.imef.org.mx
     e) Revista de la Bolsa Mexicana de Valores (BMV): www.bmv.com.mx
     f) Las instituciones de educación superior también son una fuente importante de datos —publicados o no— tanto de fuentes primarias como secundarias.

   Por supuesto, en la actualidad se puede tener acceso a gran cantidad de información producida por las organizaciones mencionadas, otra enorme cantidad de publicaciones y de fuentes de datos extranjeros u otras a través de la ya omnipresente internet.

2. Organización de datos estadísticos. Se refiere, entre otros aspectos, a la verificación de la veracidad de los datos y a su ordenación de acuerdo con las necesidades del estudio en cuestión.
3. Presentación de datos estadísticos. Se hace principalmente a través de 2 medios: tablas y gráficas. Estos medios de presentación son de suma importancia en estadística, por lo cual se revisan con detalle en el capítulo 2.
4. Análisis de datos estadísticos. Se lleva a cabo utilizando las técnicas descritas en la sección 1.1. Aquí es importante hacer notar que al planear el estudio debe decidirse cuál o cuáles técnicas estadísticas se utilizarán. Es necesario determinar la técnica estadística que puede aplicarse a la situación a analizar. No se trata de decidir "¿qué problema resuelvo con mi técnica estadística?" pues, aunque la estadística es aplicable a una gran cantidad de situaciones, no es aplicable a todas ellas; tomado esto en cuenta, debemos preguntar "¿cuál técnica estadística (o de otra área) conviene utilizar en este caso?"
5. Conclusiones. Cualquier estudio debe culminar con ellas porque, por supuesto, no se inicia un estudio con la intención de no terminarlo. Para llegar a las conclusiones deben tenerse en cuenta tanto la clase y la fuente de los datos utilizados como las técnicas aplicadas en el análisis. Es también importante cuidar la forma (y no sólo el fondo) en la presentación de las conclusiones, ya que el propósito principal de estas actividades consiste en obtener información útil para la toma de decisiones.

1.5 Datos estadísticos: variables y su clasificación

La obtención de datos para el análisis estadístico implica la medición de determinadas características: los ingresos por familia, el número de artículos defectuosos, el saldo de las cuentas por cobrar, el ingreso *per cápita*, el importe de las exportaciones, etc., y los valores que se obtienen al hacer las mediciones suelen ser diferentes, es decir, son variables.

Suelen ser diferentes: los ingresos de familias distintas, el número de artículos defectuosos en turnos o días diversos, el saldo de las cuentas por cobrar de clientes distintos, el ingreso *per cápita* de diferentes países, el importe de las exportaciones en meses diferentes, etc. Por ello suele denominarse *variable de interés,* o simplemente *variable,* a la medición que interesa en un estudio específico. En otras palabras y en términos sencillos, el término **variable** puede definirse como la característica que se mide al hacer determinadas observaciones. Es fácil ver que la denominación de *variable* refleja los diversos valores que pueden obtenerse al hacer la medición. Así, al estudiar la aceptación que un producto tiene entre los consumidores, alguna de las variables de interés sería qué tan agradable les resulta a aquéllos el sabor del producto. En este caso se tendría una variable de posibles valores como: "mucho", "poco" o "nada". De esta manera la variable sólo podría asumir estos 3 valores si así se definieran las respuestas, ya que a esta misma pregunta de qué tanto le gusta a una persona el sabor de un producto, también se le podría asociar otro conjunto distinto de categorías como: "nada", "casi nada", "un poco", "mucho" o "muchísimo".

Variable. Característica que se mide al hacer determinadas observaciones.

En otro estudio la variable de interés podría ser el ingreso mensual por familia en determinada zona, y es evidente que el número posible de valores que la variable puede asumir es muy elevado; de hecho, es infinito.

En los ejemplos mencionados puede observarse la diferencia existente entre, por un lado, la variable y, por el otro, los valores que puede tomar. En el primer ejemplo la variable es el grado de aceptación del producto y sus posibles valores son las categorías "poco", "mucho" y "nada"; mientras que en el segundo caso la variable que se mide son números, los ingresos familiares mensuales y sus posibles valores son la gran cantidad de cifras dentro de las cuales podría caer el ingreso familiar.

Estos diferentes tipos de variables, que conducen a diversas mediciones, hacen necesario revisar con más detalle los tipos de variables que aparecen en estudios estadísticos y que, a su vez, llevan a 2 tipos de clasificaciones de las variables. Por un lado, como se vio en los ejemplos anteriores, se pueden tener variables en números y variables en categorías y esto da una primera clasificación para las variables: cuantitativas y cualitativas. Sin embargo, estas 2 clases de variables se pueden subdividir aún más para dar 2 de tipo cualitativo (ordinales y nominales) y 2 de tipo cuantitativo (de intervalo y de razón). Este conjunto de 4 categorías se revisará en la sección siguiente y se agrupará bajo el título de "Escalas de medición" porque es ésta la que hace la diferencia entre una categoría y otra.

La otra clasificación de variables las divide en continuas y discontinuas (también conocidas como *discretas*), es una diferenciación muy importante en estadística y se analizará en la sección 1.7.

1.6 Escalas de medición

Puede resumirse en un cuadro la clasificación de las variables cualitativas y cuantitativas mencionada en la sección anterior:

Tabla 1.1 Variables y escalas de medición

Tipos de variables	Escala de medición
Cualitativas	Nominal
	Ordinal
Cuantitativas	De intervalo
	De razón

Este conjunto de 4 escalas de medición se denomina así porque cada nivel tiene diferente potencia o capacidad.

Escala nominal. La medición consiste en determinar si los casos específicos pertenecen a cierta categoría o no.

En la **escala nominal** la medición consiste en determinar si los casos específicos pertenecen a cierta categoría o no y, por ello, para los datos nominales sólo se pueden establecer relaciones de pertenencia o, en otras palabras, sólo se puede establecer si un caso cae en una categoría o no. Ejemplos de variables nominales son el género

(hombre o mujer), nacionalidad (mexicano, argentino, sueco, etc.), religión (budismo, judaísmo, islamismo, etc.). Vale la pena enfatizar con estos ejemplos la diferencia entre la variable y sus posibles valores:

Tabla 1.2 Variable y posibles valores

Variable	Posibles valores
Género	Hombre o mujer
Nacionalidad	Mexicano, argentino, sueco, etcétera
Religión	Budismo, judaísmo, islamismo, etcétera

Un tipo importante de datos nominales son aquellos que se dividen en sólo 2 categorías. Como ejemplo tenemos las respuestas "sí" o "no" a la pregunta "¿fuma usted?", la clasificación de productos como "defectuosos" o "no defectuosos", etc. y se trata de variables binomiales.

Por su parte, en las **escalas ordinales**, además de las relaciones de pertenencia o no pertenencia a una categoría, pueden establecerse relaciones de mayor que o después de (>) y menor que o antes de(<), ya que se conoce el orden de las categorías. Esta mayor cantidad de relaciones entre los datos hace que la escala ordinal sea más poderosa o descriptiva que la escala nominal. A continuación se presentan algunos ejemplos de variables ordinales:

Escalas ordinales. En éstas pueden establecerse relaciones de mayor que o después de (>) y menor que o antes de (<).

Tabla 1.3 Variables ordinales

Variable	Posibles valores
Rango militar	General, teniente, cabo, etcétera
Rango gerencial	Presidente, director, gerente, etcétera
Mes	Enero, febrero, marzo, etcétera
Día	Lunes, martes, miércoles, etcétera

En las escalas ordinales, la medición sólo indica la posición relativa en la escala. De ninguna manera señala la magnitud de la diferencia entre 2 posiciones distintas. Un general tiene mayor rango que un teniente pero no podría decirse qué tanto más poderoso es el primero. De la misma manera, el martes va después del lunes pero no por eso uno vale más que el otro.

En el caso de los datos numéricos (en escala de medición numérica), se distinguen 2 escalas distintas: la escala de intervalo y la escala de razón, en ese orden de potencia.

En primer lugar, la escala de intervalo es más potente que la ordinal porque con los datos en **escala de intervalo** pueden realizarse todas las operaciones aritméticas (adición, sustracción, división y multiplicación), lo cual no es válido con los datos ordinales ni tampoco con los nominales. Esta capacidad de los datos dados en escala de intervalo se debe a que, diferencias iguales en distintas partes de la escala son iguales entre sí. Por ejemplo, en la escala centígrada de temperatura, la diferencia entre cero y 10 grados centígrados de temperatura es la misma existente entre 80 y 90 grados o, en símbolos:

Escala de intervalo. En ésta pueden realizarse todas las operaciones aritméticas (adición, sustracción, división y multiplicación).

$$10° - 0° = 90° - 80°.$$

Evidentemente esta escala de intervalo es más potente que la nominal y que la ordinal.

Finalmente, la **escala de razón**, la más potente de las 4, tiene todas las capacidades de la escala de intervalo (=, ≠, >, <, +, – , ×, ÷) más una característica adicional: posee un cero absoluto que señala la carencia total de la característica que se mide.

Escala de razón. Tiene las capacidades de la escala de intervalo y además un cero absoluto que señala la carencia total de la característica que mide.

El ejemplo de la escala centígrada de temperatura como escala de intervalo implica que no se tiene un cero absoluto en el sentido de representar la carencia total del atributo: los cero grados de esta escala no representan la carencia absoluta de temperatura, ya que se sabe que el cero absoluto de temperatura se da en –273° C. El cero de la escala centígrada es relativo o, en otras palabras, arbitrario.

En cambio, en la escala de razón, el cero es absoluto: por ejemplo, al contar el número de hijos por familia, el cero representa "no hijos" o, al medir longitud, el cero representa la carencia total de esta característica.

Este cero absoluto de las escalas de razón implica una capacidad que las escalas de intervalo no tienen, ese atributo le da el nombre: permite establecer la igualdad entre 2 razones de la escala. Por ejemplo, 2/3 de centímetro son equivalentes a 10/15 de centímetro o, en símbolos:

$$\frac{2\text{ cm}}{3\text{ cm}} = \frac{10\text{ cm}}{15\text{ cm}}$$

En el caso de las escalas de intervalo no es correcto establecer estas igualdades entre 2 razones. Decir que 10°/2° es equivalente a 20°/4° no es válido, esa afirmación sería (recordando que el cero absoluto de esta escala está en los –273°):

$$\frac{-273° + 10°}{-273° + 2°} = \frac{-273° + 20°}{-273° + 4}$$

$$\frac{-263°}{-271°} = \frac{-253°}{-269°}$$

$$0.9705 = 0.9405$$

lo cual no es correcto.

Estas diferencias en la potencia de las distintas escalas permiten que las escalas superiores (más potentes) puedan convertirse a escalas inferiores (menos potentes), pero no a la inversa. Por ejemplo, puede convertirse una escala de razón en una escala de intervalo, pero no es posible convertir una escala nominal o una escala ordinal en una escala numérica. Conviene hacer hincapié en esta observación, ya que es un error cometido con demasiada frecuencia. Un ejemplo de esta práctica inadecuada es cuando las respuestas obtenidas de una encuesta en categorías como "bueno", "regular" y "malo" se convierten en números como 0, 5 y 10 para luego hacer operaciones con estos números y concluir que la calificación global es 8. Este 8 carece de sentido, ya que esta clase de conversiones no es válida porque la apreciación de las categorías es subjetiva; si a las personas les fuesen asignados números para su evaluación, lo más probable es que la calificación global que se obtuviese sería bastante diferente. Por otro lado, con el mismo ejemplo, tampoco tendría sentido pedir que las opiniones se dieran directamente en números, ya que lo que se indaga son precisamente opiniones: el 7 puede significar regular para una persona y malo para otra. Además, tampoco es necesario realizar estas conversiones equivocadas, ya que existen métodos estadísticos para el análisis de este tipo de datos.

Finalmente, aunque es posible la conversión de una escala superior a una inferior, por lo general conviene evitar hacer este tipo de manipulaciones pues al hacerlo se pierde información (potencia).

1.7 Variables continuas y variables discontinuas (o discretas)

Datos continuos. Se pueden expresar con tal precisión que llega un momento en el que es difícil distinguir entre un número y el siguiente.

Esta importante distinción de las variables es considerablemente sencilla: los **datos continuos** son los que pueden expresarse con tal precisión que llega un momento en el que es difícil distinguir entre un número y el siguiente. Por ejemplo, el peso de un animal se puede expresar en kilogramos, en gramos, en miligramos o, en breve, en unidades tan pequeñas que distinguir entre 50.00000000001 y 50.00000000002 kilogramos es prácticamente imposible. Por ello se les llama *continuos*: puede no haber divisiones o espacios entre cualquiera de ellos con el que le antecede o le sucede. En general, las mediciones de dimensiones como peso y longitud dan lugar a datos de este tipo.

Datos discontinuos o **discretos**. Por su naturaleza, se expresan en cantidades fácilmente distinguibles unas de otras.

Los **datos discontinuos** (también conocidos como *discretos*) son los que por su naturaleza se expresan en cantidades fácilmente distinguibles unas de otras. Por ejemplo, el número de hijos, la cantidad de artículos defectuosos y el número de errores mecanográficos por hoja, son datos discretos. Conviene observar que tanto los datos nominales como los jerarquizados son mediciones de tipo discreto, mientras que las mediciones con números pueden ser discretas o continuas.

Otra manera de ilustrar la continuidad y la discontinuidad entre datos es mediante la comparación con el trazado de una línea sobre papel, sin levantar el lápiz sería de continuidad, en tanto que una línea discreta o discontinua estaría representada por una línea punteada trazada levantando repetidamente el lápiz del papel para dejar huecos.

A lo largo de todo el libro se mencionará, según resulte pertinente, la clase de datos con los que se trabaja.

1.8 Uso de computadoras en estadística

Es muy conveniente el uso de computadoras en estadística, ya que se manejan considerables cantidades de datos. La tecnología sirve para hacer que el trabajo sea más rápido y más eficiente. Sin embargo, la

apropiada utilización de las computadoras exige que quien las emplee sepa cuáles son sus capacidades y sus limitaciones; es ya un lugar común la sentencia de que si se introduce basura (datos o instrucciones deficientes) en la computadora, ésta producirá precisamente basura (resultados inútiles y engañosos), *garbage in*, *garbage out* o GIGO por sus siglas en inglés. Por ello, en el presente texto primeramente se hará hincapié en la comprensión de conceptos y de los procedimientos de cálculo, antes de utilizar un paquete de computación para facilitar el trabajo.

Son muchos los paquetes o plataformas de computación comerciales especialmente diseñados para el manejo estadístico de grandes o pequeñas cantidades de datos. Entre los más comunes está el *Statistical Package for the Social Sciences* (SPSS, o Paquete Estadístico para las Ciencias Sociales) y el Minitab. Estos paquetes son muy potentes y completos, permiten realizar estudios estadísticos desde sencillos hasta muy complejos.

Por otro lado, existen numerosos sitios de internet, sobre todo en inglés, que ofrecen paquetes estadísticos gratuitos. Una búsqueda simple suele arrojar numerosos resultados.

Sin embargo, para quienes estudian un curso básico de estadística como el que se cubre en el presente texto —cuyo propósito principal es que el estudiante comprenda las técnicas estadísticas para que sea capaz de aplicarlas en la práctica— resulta más sencillo y también muy útil emplear una herramienta como la hoja de cálculo Excel de Microsoft, una herramienta tan popular que la mayoría de los estudiantes y profesionales ejecutan prácticamente todos los días; Excel también ofrece una amplia variedad de aplicaciones estadísticas.

Por lo anterior, en este libro se incluyen abundantes ejemplos de cómo utilizar Excel para facilitar el trabajo, una vez que el estudiante tenga una comprensión adecuada de los conceptos y los procedimientos de cálculo. Se manejará en este libro la versión 2007 de este paquete; en la sección siguiente se resumirán sus principales capacidades estadísticas.

1.9 Excel 2007

La organización y presentación de datos en forma de gráficas es una herramienta estadística muy útil de Excel, es posible que una de las principales ayudas que ofrece en el campo de la estadística sea la versátil y sencilla manera en que los usuarios pueden hacer gráficas. Pero como la elaboración de gráficas con calidad profesional depende principalmente del usuario, en el capítulo 2 se utilizarán numerosos ejemplos con Excel para revisar los elementos y los principios en la elaboración de gráficas.

Además de las evidentes capacidades para organizar datos, como ordenar un conjunto de números o contar celdas que tengan entradas con determinadas características, Excel 2007 (en lo sucesivo simplemente Excel) cuenta con un numeroso conjunto de funciones estadísticas. En el apéndice 1 se presenta un glosario de todas ellas, con su nombre en Excel y con una breve descripción de lo que producen y que, por supuesto y con mayor detalle, también pueden consultarse en la sección de "Ayuda" del propio software. A lo largo del texto se presentarán ejemplos de cómo resolver aplicaciones estadísticas utilizando estas funciones.

Asimismo, Excel cuenta con un grupo de herramientas para un análisis estadístico considerablemente más completo que el elaborado por las funciones simples. Este conjunto de herramientas se encuentra en la pestaña "Datos" del menú principal; el usuario debe saber que esta función no siempre se activa automáticamente al instalar Excel. Si no aparece ahí, es necesario descargarla desde el menú "Ayuda"; para ello se busca la opción "Análisis de datos" y con seguir las sencillas instrucciones que ahí se detallan queda disponible. Las herramientas que tiene son las siguientes y entre paréntesis se señala el número del capítulo en donde se tratará cada una de ellas:

- Análisis de varianza de un factor (12).
- Análisis de varianza de 2 factores con varias muestras por grupo (12).
- Análisis de varianza de 2 factores con una sola muestra por grupo (12).
- Coeficiente de correlación (13 y 14).
- Covarianza (13 y 14).
- Estadística descriptiva (3)
- Suavización exponencial (15).
- Prueba F para varianzas de dos muestras (10).
- Análisis de Fourier (no se revisa en este texto).
- Histograma (2).
- Media móvil (15).

- Generación de números aleatorios (7).
- Jerarquía y percentil (3).
- Regresión (13 y 14).
- Muestra (7).
- Prueba t para medias de 2 muestras emparejadas (10).
- Prueba t para 2 muestras suponiendo varianzas iguales (10).
- Prueba t para 2 muestras suponiendo varianzas desiguales (10).
- Prueba Z para medias de 2 muestras (10).

Al igual que con las funciones, conforme sea conveniente, se resolverán ejemplos para ilustrar la aplicación de estas herramientas.

1.10 Resumen

La estadística es la parte de las matemáticas que, siguiendo su propio método y aplicando diversas técnicas, permite el análisis de varios tipos de datos con el propósito de obtener conclusiones e información que asistan en el proceso de toma de decisiones. El propósito del libro es revisar las principales técnicas de la estadística aplicada a datos relevantes para los negocios y la economía.

La división de la estadística en 2 grandes áreas (estadística descriptiva y estadística inferencial) señala sus 2 grandes campos de acción, y a su vez explica el desarrollo que esta disciplina ha tenido: fue primordialmente descriptiva y avanzó en sus capacidades de inferencia hasta llegar al punto actual, en el que la inferencia estadística es un campo de muy amplia aplicación y en pleno crecimiento.

En estadística se utiliza con frecuencia el término *variable* para hacer referencia a la característica que se mide, como por ejemplo sueldo, edad o género. Es importante saber que las variables estadísticas pueden medirse en 4 escalas: 2 cualitativas (nominal y ordinal) y 2 cuantitativas (de intervalo y de razón) y además pueden ser continuas o discontinuas.

Aparte de paquetes de computación comerciales y gratuitos para el análisis estadístico, su difundido uso y grandes capacidades hacen del programa Excel de Microsoft una herramienta muy útil para facilitar el análisis estadístico. En este texto se utiliza la versión 2007 para resolver numerosos ejemplos de aplicación una vez que el estudiante comprende los conceptos y practica manualmente los procedimientos de cálculo, antes de utilizar Excel para facilitar el trabajo.

EJERCICIOS 1.1 Ajuste por discontinuidad

1. ¿Qué es estadística?

2. Explique las diferencias entre:

a) Población y muestra.
b) Parámetros y estadísticos.
c) Estadística descriptiva y estadística inferencial.
d) Variables discretas y variables continuas.

3. Explique las diferentes escalas de medición.

4. Anote 3 definiciones de estadística que aparezcan en libros similares a éste.

5. En una hoja tabular, dividida en 3 columnas, anote en cada una de éstas el listado de contenido de 3 libros de estadística, incluyendo éste.

a) ¿Qué similitudes encuentra en los 3?
b) ¿Cuáles diferencias importantes encuentra?

6. ¿Qué tienen que ver los datos estadísticos con mediciones?

7. Para los siguientes datos diga si se trata de: *a*) datos discretos o continuos y *b*) la escala de medición.

a) Las edades de un grupo de personas.
b) La longitud de las piezas que fabrica una máquina.
c) La antigüedad en el puesto de los empleados de una compañía.
d) La preferencia por partidos políticos entre ciudadanos.
e) El número de artículos defectuosos en un lote de producción.
f) Las calificaciones de un grupo de empleados que participaron en un curso de capacitación y que se especifican como S (satisfactorio), B (bueno) y M (magnífico).

8. ¿Qué tienen que ver las computadoras con la estadística?

9. Haga un breve resumen de la historia de la estadística.

10. Describa qué clase de información se puede obtener de la sección "Otras fuentes oficiales de información estadística" que aparece en la sección de "Estadísticas" de la página de internet del Banco de México (consultada el 5 de abril de 2011): www.banxico.gob.mx/estadisticas/index.html.

11. ¿Puede utilizar en su computadora la sección de "Análisis de Datos" de la pestaña "Datos" de su programa de Microsoft Excel? Explórela.

capítulo 2

Presentación de datos: tablas y gráficas

Sumario

2.1 Principales elementos de tablas y gráficas
2.2 Tablas
- 2.2.1 Series simples, series de datos y frecuencias, y series de clases y frecuencias
- 2.2.2 Tablas de frecuencias para datos cualitativos
- 2.2.3 Frecuencias absolutas, relativas y acumuladas
- 2.2.4 Tablas de doble entrada o de clasificación cruzada o de contingencias
- 2.2.5 Uso de Excel. Construcción de distribuciones de frecuencias

2.3 Gráficas (con Excel)
- 2.3.1 Histogramas
- 2.3.2 Gráficas de líneas
- 2.3.3 Histogramas y polígonos de frecuencias
- 2.3.4 Gráficas circulares
- 2.3.5 Otras aplicaciones

2.4 Resumen
2.5 Ejercicios adicionales

La forma más conveniente de resumir conjuntos numerosos de datos es a través de tablas y gráficas, ya que permiten condensar la información y, al mismo tiempo, facilitan la apreciación de su contenido. Por lo general, para construir gráficas es necesario resumir primero los datos en una tabla y, por ello, la relación entre estas 2 formas de presentación de datos es estrecha. Sin embargo, hay ocasiones que en la presentación final se prefiera alguna de las 2 formas: cuando se desea transmitir una idea rápida y sencilla se utiliza una gráfica; mientras que cuando es necesario analizar los datos con mayor detenimiento o precisión, las tablas son más útiles. Por supuesto, existen casos en los que conviene utilizar ambas formas, junto con algunos comentarios adicionales.

Por otro lado, es posible construir una variedad enorme de tablas y gráficas, por lo que en las secciones siguientes se presentan sólo los tipos más comunes y de mayor utilidad.

Para elaborar buenas tablas y gráficas conviene tomar en consideración diversos aspectos, tales como el propósito de la tabla (a quién va dirigida, cuál es la información importante que no debe pasar desapercibida, etc.); la clase de datos que se quiere presentar; colocación de éstos para facilitar el análisis, y otros.

2.1 Principales elementos de tablas y gráficas

Ambas formas de resumen y presentación de datos comparten características que les son comunes. En general, las tablas y gráficas que se usan para presentaciones o escritos formales tienen los siguientes elementos:

Tablas	Gráficas
Título o encabezado	Título o encabezado
Subtítulo	Subtítulo
Cuerpo	Figura o gráfica
Fuente de los datos	Fuente de los datos
Encabezados	Leyendas
Notas	Notas

Por lo regular, no todas las gráficas y tablas requieren de todos estos elementos, pero el título, el cuerpo, la figura o gráfica, y posiblemente la fuente de datos deben considerarse necesarios. En lo que se

refiere a tablas, en la sección siguiente se analizan, por un lado, 3 tipos de tablas que pueden construirse a partir de un conjunto determinado de datos (series simples, series de datos y frecuencias, y series de clases y frecuencias), son de especial interés en estadística. Por otra parte, existen las tablas de doble entrada que también son de particular interés. Por último, respecto a las gráficas, en la sección 2.3 se revisarán los principales tipos de gráficas: de barras e histogramas, de círculo o de pastel y de línea.

2.2 Tablas

Tal como se mencionó, existe una amplia variedad de tablas, ya que es posible elaborarlas de diversas maneras, colocando la información en muchas formas distintas e incluyendo varios elementos según sea necesario o conveniente. Sin embargo, uno de los principales criterios de clasificación para propósitos estadísticos es el que se refiere a cuántas variables contiene la tabla. De acuerdo con este criterio se tendrían tablas de una sola entrada (una sola variable), de doble entrada (2 variables) y de entradas múltiples (más de 2 variables).

Series. Conjuntos de datos que se presentan en tablas.
Series de datos agrupados. Tablas de datos en las que se resumen éstos de acuerdo con la frecuencia con la que se repiten o según determinados intervalos de valores.

Distribuciones de frecuencias. Series que utilizan frecuencias.

A los conjuntos de datos que se presentan en tablas se les llaman **series**, de manera que puede hablarse de series simples y series de datos agrupados. Estas **series de datos agrupados** son tablas de datos en las que se resumen éstos de acuerdo con la frecuencia con la que se repiten o según determinados intervalos de valores. En otras palabras, a estas series de datos agrupados se les puede dividir, a su vez, en series de datos y frecuencias, y series de clases y frecuencias. En los ejemplos que se presentan más adelante se clarificarán estos conceptos.

Además, tal como se verá más adelante, a las series que utilizan frecuencias se les conoce también como **distribuciones de frecuencias** porque las frecuencias permiten apreciar la forma en la que se distribuyen o comportan los datos (puede apreciarse, por un lado, si están cargados hacia un extremo o el otro, si están aglutinados cerca del centro, o si están distribuidos de manera uniforme).

2.2.1 Series simples, series de datos y frecuencias, y series de clases y frecuencias

Para ilustrar estos 3 tipos de series se utiliza el conjunto de datos que publica cada año la Comisión Nacional de los Salarios Mínimos (Conasami), información que trata sobre los salarios mínimos vigentes en la República Mexicana, por lo general durante cada año.

EJEMPLO 2.1

La Conasami estableció en 2011 lo que denomina como salarios mínimos generales, además de un conjunto de salarios mínimos profesionales para 72 profesiones, oficios y trabajos especiales para cada una de 3 áreas geográficas. En la tabla 2.1 se muestran los salarios mínimos por profesión que entraron en vigor el 1 de enero de 2011.

Tabla 2.1 Salarios mínimos vigentes en México a partir del 1 de enero de 2011 (en pesos)

		Áreas geográficas		
Oficio núm.	**Profesiones, oficios y trabajos especiales**	**A**	**B**	**C**
1	Albañilería, oficial de	87.17	84.92	82.63
2	Boticas, farmacias y droguerías, dependiente de mostrador en	75.86	73.84	71.97
3	*Bulldozer* y/o traxcavo, operador de	91.83	89.21	86.89
4	Cajero(a) de máquina registradora	77.33	75.43	73.52
5	Cajista de imprenta, oficial	82.31	80.25	77.95
6	Cantinero preparador de bebidas	79.13	76.98	74.98
7	Carpintero en fabricación y reparación de muebles, oficial	85.57	83.21	80.98
8	Cocinero(a), mayor(a) en restaurantes, fondas y demás establecimientos de preparación y venta de alimentos	88.45	86.15	83.74
9	Colchones, oficial en fabricación y reparación de	80.01	77.94	76
10	Colocador de mosaicos y azulejos, oficial	85.2	83.06	80.78

		Áreas geográficas		
Oficio núm.	**Profesiones, oficios y trabajos especiales**	**A**	**B**	**C**
11	Construcción de edificios y casas habitación, yesero en	80.65	78.71	76.47
12	Cortador en talleres y fábricas de manufactura de calzado, oficial	78.28	76.34	74.34
13	Costurero(a) en confección de ropa en talleres o fábricas	77.2	74.98	73.32
14	Costurero(a) en confección de ropa en trabajo a domicilio	79.5	77.48	75.31
15	Chofer acomodador de automóviles en estacionamientos	81.28	79.1	76.91
16	Chofer de camión de carga en general	89.2	86.91	84.64
17	Chofer de camioneta de carga en general	86.39	84.09	81.74
18	Chofer operador de vehículos con grúa	82.69	80.7	78.46
19	Draga, operador de	92.78	90.44	87.86
20	Ebanista en fabricación y reparación de muebles, oficial	86.97	84.66	82.32
21	Electricista instalador y reparador de instalaciones eléctricas, oficial	85.2	83.06	80.78
22	Electricista en la reparación de automóviles y camiones, oficial	86.14	83.84	81.49
23	Electricista reparador de motores y/o generadores en talleres de servicio, oficial	82.69	80.7	78.46
24	Empleado de góndola, anaquel o sección en tiendas de autoservicio	75.61	73.64	71.32
25	Encargado de bodega y/o almacén	78.68	76.59	74.59
26	Enfermería, auxiliar práctico de	81.28	79.1	76.91
27	Ferreterías y tlapalerías, dependiente de mostrador en	80.45	78.13	76.15
28	Fogonero de calderas de vapor	83.33	81.08	78.85
29	Gasolinero, oficial	77.2	74.98	73.32
30	Herrería, oficial de	83.97	81.78	79.48
31	Hojalatero en la reparación de automóviles y camiones, oficial	85.57	83.21	80.98
32	Hornero fundidor de metales, oficial	87.67	85.51	83.23
33	Joyero-platero, oficial	81.28	79.1	76.91
34	Joyero-platero en trabajo a domicilio, oficial	84.67	82.62	80.33
35	Linotipista, oficial	90.38	88.19	85.87
36	Lubricador de automóviles, camiones y otros vehículos de motor	77.9	75.84	73.68
37	Maestro en escuelas primarias particulares	92.14	89.8	87.23
38	Manejador en granja avícola	74.65	72.8	70.86
39	Maquinaria agrícola, operador de	87.67	85.51	83.23
40	Máquinas para madera en general, oficial operador de	83.33	81.08	78.85
41	Mecánico en reparación de automóviles y camiones, oficial	90.38	88.19	85.87
42	Mecánico tornero, oficial	84.67	82.62	80.33
43	Moldero en fundición de metales	82.69	80.7	78.46
44	Montador en talleres y fábricas de calzado, oficial	78.28	76.34	74.34
45	Niquelado y cromado de artículos y piezas de metal, oficial de	82.31	80.25	77.95
46	Peinador(a) y manicurista	81.28	79.1	76.91
47	Perforista con pistola de aire	86.14	83.84	81.49
48	Pintor de automóviles y camiones, oficial	83.97	81.78	79.48
49	Pintor de casas, edificios y construcciones en general, oficial	83.33	81.08	78.85
50	Planchador a máquina en tintorerías, lavanderías y establecimientos similares	77.33	75.43	73.52
51	Plomero en instalaciones sanitarias, oficial	83.51	81.47	79.17
52	Prensa *offset* multicolor, operador de	87.17	84.92	82.63
53	Prensista, oficial	81.28	79.1	76.91
54	Radiotécnico reparador de aparatos eléctricos y electrónicos, oficial	86.97	84.66	82.32
55	Recamarero(a) en hoteles, moteles y otros establecimientos de hospedaje	75.61	73.64	71.32
56	Refaccionarias de automóviles y camiones, dependiente de mostrador en	78.68	76.59	74.59

(*continúa*)

Tabla 2.1 (*continuación*)

		Áreas geográficas		
Oficio núm.	**Profesiones, oficios y trabajos especiales**	**A**	**B**	**C**
57	Reparador de aparatos eléctricos para el hogar, oficial	82.31	80.25	77.95
58	Reportero(a) en prensa diaria impresa	179.2	174.67	169.6
59	Reportero(a) gráfico(a) en prensa diaria impresa	179.2	174.67	169.6
60	Repostero o pastelero	87.17	84.92	82.63
61	Sastrería en trabajo a domicilio, oficial de	87.67	85.51	83.23
62	Secretario(a) auxiliar	90.2	87.67	85.51
63	Soldador con soplete o con arco eléctrico	86.14	83.84	81.49
64	Talabartero en la manufactura y reparación de artículos de piel, oficial	81.28	79.1	76.91
65	Tablajero y/o carnicero en mostrador	81.28	79.1	76.91
66	Tapicero de vestiduras de automóviles, oficial	82.69	80.7	78.46
67	Tapicero en reparación de muebles, oficial	82.69	80.7	78.46
68	Trabajo social, técnico(a) en	98.6	95.88	93.39
69	Vaquero ordeñador a máquina	75.61	73.64	71.32
70	Velador	77.2	74.98	73.32
71	Vendedor de piso de aparatos de uso doméstico	79.5	77.48	75.31
72	Zapatero en talleres de reparación de calzado, oficial	78.28	76.34	74.34

Fuente: Servicio de Administración Tributaria, *Salarios mínimos 2012*, disponible en: http://www.sat.gob.mx/sitio_internet/asistencia_contribuyente/informacion_frecuente/salarios_minimos/, consultado el 31 de marzo de 2011.

Como puede verse en la tabla, existen 72 categorías de salarios profesionales y México está dividido en 3 áreas geográficas, de acuerdo al costo de vida. Puede apreciarse que los salarios mínimos de la zona A son más altos porque es la zona con más elevado costo de vida, seguida por las zonas B y C en ese orden.

La tabla 2.1 contiene 3 series simples de 72 salarios mínimos, uno para cada una de las 3 zonas geográficas. En otras palabras, una sola serie simple sería el conjunto de los 72 salarios profesionales de la zona A; otra representaría los de la zona B, y la tercera estaría constituida por los de la zona C.

2.2.1.1 Series simples

Si se utilizan por el momento sólo los salarios de la zona A se tendría entonces la serie simple que aparece en la tabla 2.2 que además está ordenada de menor a mayor salario.

Tabla 2.2 Serie simple de los salarios mínimos vigentes durante 2011 en el área A (ordenados de menor a mayor)

Profesiones, oficios y trabajos especiales	Área geográfica A
Manejador en granja avícola	74.65
Empleado de góndola, anaquel o sección en tiendas de autoservicio	75.61
Recamarero(a) en hoteles, moteles y otros establecimientos de hospedaje	75.61
Vaquero ordeñador a máquina	75.61
Boticas, farmacias y droguerías, dependiente de mostrador en	75.86
Costurero(a) en confección de ropa en talleres o fábricas	77.2
Gasolinero, oficial	77.2
Velador	77.2
Cajero(a) de máquina registradora	77.33
Planchador a máquina en tintorerías, lavanderías y establecimientos similares	77.33
Lubricador de automóviles, camiones y otros vehículos de motor	77.9
Cortador en talleres y fábricas de manufactura de calzado, oficial	78.28
Montador en talleres y fábricas de calzado, oficial	78.28
Zapatero en talleres de reparación de calzado, oficial	78.28
Encargado de bodega y/o almacén	78.68
Refaccionarias de automóviles y camiones, dependiente de mostrador en	78.68

Profesiones, oficios y trabajos especiales	Área geográfica A
Cantinero preparador de bebidas	79.13
Costurero(a) en confección de ropa en trabajo a domicilio	79.5
Vendedor de piso de aparatos de uso doméstico	79.5
Colchones, oficial en fabricación y reparación de	80.01
Ferreterías y tlapalerías, dependiente de mostrador en	80.45
Construcción de edificios y casas habitación, yesero en	80.65
Chofer acomodador de automóviles en estacionamientos	81.28
Enfermería, auxiliar práctico de	81.28
Joyero-platero, oficial	81.28
Peinador(a) y manicurista	81.28
Prensista, oficial	81.28
Talabartero en la manufactura y reparación de artículos de piel, oficial	81.28
Tablajero y/o carnicero en mostrador	81.28
Cajista de imprenta, oficial	82.31
Niquelado y cromado de artículos y piezas de metal, oficial de	82.31
Reparador de aparatos eléctricos para el hogar, oficial	82.31
Chofer operador de vehículos con grúa	82.69
Electricista reparador de motores y/o generadores en talleres de servicio, oficial	82.69
Moldero en fundición de metales	82.69
Tapicero de vestiduras de automóviles, oficial	82.69
Tapicero en reparación de muebles, oficial	82.69
Fogonero de calderas de vapor	83.33
Máquinas para madera en general, oficial operador de	83.33
Pintor de casas, edificios y construcciones en general, oficial	83.33
Plomero en instalaciones sanitarias, oficial	83.51
Herrería, oficial de	83.97
Pintor de automóviles y camiones, oficial	83.97
Joyero-platero en trabajo a domicilio, oficial	84.67
Mecánico tornero, oficial	84.67
Colocador de mosaicos y azulejos, oficial	85.2
Electricista instalador y reparador de instalaciones eléctricas, oficial	85.2
Carpintero en fabricación y reparación de muebles, oficial	85.57
Hojalatero en la reparación de automóviles y camiones, oficial	85.57
Electricista en la reparación de automóviles y camiones, oficial	86.14
Perforista con pistola de aire	86.14
Soldador con soplete o con arco eléctrico	86.14
Chofer de camioneta de carga en general	86.39
Ebanista en fabricación y reparación de muebles, oficial	86.97
Radiotécnico reparador de aparatos eléctricos y electrónicos, oficial	86.97
Albañilería, oficial de	87.17
Prensa *offset* multicolor, operador de	87.17
Repostero o pastelero	87.17
Hornero fundidor de metales, oficial	87.67
Maquinaria agrícola, operador de	87.67
Sastrería en trabajo a domicilio, oficial de	87.67
Cocinero(a) mayor(a) en restaurantes, fondas y demás establecimientos de preparación y venta de alimentos	88.45
Chofer de camión de carga en general	89.2
Secretario(a) auxiliar	90.2
Linotipista, oficial	90.38

(*continúa*)

Tabla 2.2 (*continuación*)

Profesiones, oficios y trabajos especiales	Área geográfica A
Mecánico en reparación de automóviles y camiones, oficial	90.38
Bulldozer y/o traxcavo, operador de	91.83
Maestro en escuelas primarias particulares	92.14
Draga, operador de	92.78
Trabajo social, técnico(a) en	98.6
Reportero(a) en prensa diaria impresa	179.2
Reportero(a) gráfico(a) en prensa diaria impresa	179.2

Esta serie simple ordenada ofrece cierta información adicional: el oficio con el menor salario mínimo es el de manejador en granja avícola, y el que tiene el mayor salario es el de reportero(a) gráfico(a) en prensa diaria impresa. Como los datos de la serie simple anterior son numerosos, ya que está conformada por 72 oficios, conviene compactarlos para apreciar mejor la composición de los salarios. Para ello, en la tabla 2.3 se les agrupa en una serie de datos y frecuencias que utiliza la serie ordenada de la tabla 2.2, y cuenta las veces que aparece cada salario.

2.2.1.2 Series de datos y frecuencias

Este tipo de tablas presenta una ventaja considerable sobre las tablas simples. En las tablas de frecuencias se agrupan los datos de manera que si existen observaciones repetidas (datos con el mismo valor) no se enumera cada una de ellas sino que se menciona una sola vez el valor repetido, y a éste se le asocia el número de repeticiones (frecuencia de aparición). El nombre de estas tablas se debe a que en una columna se identifican los datos, los valores que toma la variable (x), y en una segunda columna se especifica la frecuencia (f) con la que aparece cada valor. Obsérvese que esa x que se usa como encabezado de los valores de la variable es la forma común de identificar la columna que contiene los datos, mientras que la f identifica la frecuencia de aparición de cada uno de aquéllos.

El procedimiento consiste en determinar cuántas veces se repite cada salario; para hacer esto se anotan en la primera columna los salarios diferentes, como la que aparece en la tabla 2.3 y en la segunda se registran las veces que se repite cada salario:

Tabla 2.3 Salarios mínimos diferentes de la zona geográfica A

Salarios de la zona A x	Frecuencia f
74.65	1
75.61	3
75.86	1
77.2	3
77.33	2
77.9	1
78.28	3
78.68	2
79.13	1
79.5	2
80.01	1
80.45	1
80.65	1
81.28	7
82.31	3
82.69	5
83.33	3
83.51	1
83.97	2

Salarios de la zona A x	Frecuencia f
84.67	2
85.2	2
85.57	2
86.14	3
86.39	1
86.97	2
87.17	3
87.67	3
88.45	1
89.2	1
90.2	1
90.38	2
91.83	1
92.14	1
92.78	1
98.6	1
179.2	2
Suma de las frecuencias	72

Nótese cómo la suma de las frecuencias es precisamente el número de salarios de todos los trabajadores considerados en la tabla original. Además, la variable "salarios mínimos de la zona A" se representa mediante x y la frecuencia mediante f.

La principal ventaja de esta distribución de frecuencias con respecto a la serie simple, la original, es que permite presentar los datos en forma más resumida, más compacta. La serie simple tenía 72 renglones más el encabezado y esta serie de datos y frecuencias tiene 38, contando el encabezado y el último renglón en donde se asienta la suma de las frecuencias; sin embargo, aún parecen ser demasiados renglones. En la distribución de clases y frecuencias que se ilustrará en la sección siguiente esta compactación de los datos es aún mayor.

Tabla 2.4 Serie de clases y frecuencias con los salarios mínimos vigentes en 2011 en el área A

Clases de salarios x	Frecuencia f
70 a menos de 80	19
80 a menos de 90	44
90 a menos de 100	7
100 a menos de 110	0
110 a menos de 120	0
120 a menos de 130	0
130 a menos de 140	0
140 a menos de 150	0
150 a menos de 160	0
160 a menos de 170	0
170 a menos de 180	2
	72

2.2.1.3 Series de clases y frecuencias

En la serie anterior se estableció la variable x para cada uno de los distintos valores de los salarios mínimos de la zona geográfica A. Ahora se procede a establecer x, no por valores únicos, sino por intervalos o clases de valores.

Antes se revisó que los valores máximo y mínimo de los salarios de la zona A son 179.2 y 74.65, respectivamente. Si se resta el segundo al primero, 179.2 – 74.65 = 104.55, se obtiene el rango dentro del que se encuentran todos los salarios. Si se redondea este rango a 110, existe la posibilidad de dividir ese rango de valores de aproximadamente 110 unidades que van de un valor redondo, inferior al mínimo, 70, hasta otro valor redondo superior al máximo, 180 en 10 intervalos de 10 unidades, de la siguiente manera:

Variable x salarios mínimos
70 a menos de 80
80 a menos de 90
90 a menos de 100
100 a menos de 110
110 a menos de 120
120 a menos de 130
130 a menos de 140
140 a menos de 150
150 a menos de 160
160 a menos de 170
170 a menos de 180

Al utilizar la serie de datos y frecuencias que se construyó previamente en la tabla 2.3 puede construirse, contando cuántos de los datos caen dentro de cada intervalo, la tabla 2.4, que será la serie de clases y frecuencias, y resume aún más los datos originales de los salarios mínimos profesionales del área A.

Tal como se esperaba, esta distribución de frecuencias agrupadas en clases es mucho más compacta que las 2 anteriores, la simple y la de datos y frecuencias.

Tabla 2.5 Distribución de frecuencias agrupadas de los salarios mínimos vigentes en 2011 en la zona B

Clases de salarios x	Frecuencia f
70 a menos de 80	29
80 a menos de 90	39
90 a menos de 100	2
100 a menos de 110	0
110 a menos de 120	0
120 a menos de 130	0
130 a menos de 140	0
140 a menos de 150	0
150 a menos de 160	0
160 a menos de 170	0
170 a menos de 180	2
Total	72

Es muy importante tener un detalle en cuenta sobre estos 3 tipos de tablas en las que se presentaron los datos de los salarios mínimos, se trata de que mientras que la tabla simple y la de datos y frecuencias contienen exactamente los mismos datos (en la primera sin agrupar y en la segunda agrupados), en la tabla de clases y frecuencias se pierde un poco de información y no se puede reconstruir si no se cuenta con los datos originales. El primer intervalo, "70 a menos de 80", tiene 19 observaciones; en otras palabras, hay 19 datos que están entre los valores de 70 y 80 pero si no se contara con los datos originales y completos de la Conasami, no habría manera de determinar el valor preciso de esos 19 datos. Esto resultaría ser un inconveniente en algunos casos pero por lo general no es un problema, sobre todo cuando se trata de condensar grandes cantidades de datos, como en los censos, por ejemplo.

Con respecto a la conversión de series simples en series de datos agrupados en formato de clases y frecuencias, tal como se ilustró, es perfectamente posible pasar de una serie simple a una de datos y frecuencias, y finalmente a una de clases y frecuencias. Sin embargo, el proceso inverso no siempre es posible; es decir, si se parte de una serie originalmente obtenida en forma de tabla de clases y frecuencias y no se tiene acceso a los datos originales, no es posible convertirla en una serie de datos y frecuencias o en una serie simple.

Siguiendo el mismo procedimiento descrito antes se obtuvieron las distribuciones de clases y frecuencias para los salarios de las zonas geográficas B y C y son las que se muestran en las tablas 2.5 y 2.6.

Tabla 2.6 Distribución de frecuencias agrupadas de los salarios mínimos vigentes en 2011 en la zona C

Clases de salarios *x*	Frecuencia *f*
70 a menos de 80	43
80 a menos de 90	26
90 a menos de 100	1
100 a menos de 110	0
110 a menos de 120	0
120 a menos de 130	0
130 a menos de 140	0
140 a menos de 150	0
150 a menos de 160	0
160 a menos de 170	2
170 a menos de 180	0
Total	72

Estas 3 tablas de frecuencias agrupadas en clases ilustran cómo este conglomerado hace posible compactar tablas de mayor tamaño y permite sacar conclusiones rápidas por medio de la observación de los datos.

Lo que salta a la vista es que en las 3 zonas geográficas los salarios se concentran en la parte baja de la escala, y una segunda conclusión es que en la zona B, de menor nivel de costo de vida que la A, los salarios están aún más concentrados en la zona de los salarios más bajos, asimismo esta tendencia es todavía mayor en la zona de menor nivel, la zona C.

2.2.1.4 Construcción de tablas de clases y frecuencias

La conversión de series simples o de datos y frecuencias en series de clases y frecuencias implica la determinación del número de clases que ésta habrá de tener y de la amplitud con la que contará cada clase o intervalo. Estas decisiones dependen de varios factores como el uso que se desea dar a los datos, el tipo de datos con los que se trabaja y hasta los propios gustos de quien hace el trabajo; por lo anterior existen diversas propuestas para determinar este valor.

En los ejemplos que se presentaron simplemente se determinaron las clases para utilizar números redondos, con clases que comenzaban en decenas, empezando por "70 a menos de 80" y continuando así hasta abarcar el valor máximo.

Regla de Sturges. Primera aproximación al número de clases que debe tener la serie de clases y frecuencias mediante la raíz cuadrada del número de elementos.

Otra propuesta de uso común es la **regla de Sturges**, que consiste en obtener una primera aproximación de cuántas clases debe tener la serie de clases y frecuencias mediante la raíz cuadrada del número de elementos. Así, como se tienen 72 datos y la raíz cuadrada de 72 es 8.49, la sugerencia inicial será utilizar 9 clases para construir las demás clases según el número de datos. De nuevo, en el ejemplo de los salarios mínimos de la zona A, se tenían como máximo y mínimo 179.2 y 74.65, lo que da una diferencia de 104.55. Y como la regla de Sturges indica 9 clases, cada una debe medir aproximadamente 104.55/9 = 11.61 que, redondeado, da 12. Se presenta en seguida un ejemplo que ilustra este procedimiento según la regla de Sturges.

EJEMPLO 2.2

La siguiente tabla contiene las edades de una muestra aleatoria de 80 empleados de una empresa. Construir una tabla de clases y frecuencias.

40	50	42	40	41	54	47	55	30	45
21	70	60	31	45	50	54	50	30	35
30	19	50	52	29	25	60	60	34	47
50	45	60	55	30	35	40	48	43	56
70	58	50	65	32	41	48	40	55	53
51	68	65	85	49	75	45	52	40	42
47	66	58	20	48	37	69	55	65	53
49	46	40	51	55	73	20	50	75	52

Solución:
En este ejemplo se tienen 80 datos, como la raíz cuadrada de 80 es 8.94, la sugerencia será utilizar 9 clases.

Una vez que se determina el número de clases se revisarán los valores máximo y mínimo de los datos para determinar el tamaño de cada clase y el límite inferior de la primera clase y el límite superior de la última.

En este ejemplo, al inspeccionar la tabla, se ve que la edad máxima es 85 y la mínima 19, por lo que el rango en el cual se encuentran todos los datos es 85 – 19 = 66; si se va a dividir este rango en 9 clases, entonces cada una de ellas tendrá una amplitud aproximada de 66/9 = 7.33. Parece que convendría redondear este número a 8, un número par.

Como el mínimo es 19, parece adecuado empezar la primera clase en el valor 18 (otro número par) y construir 9 clases de amplitud 8, para obtener el siguiente conjunto de clases:

18 a menos de 26
26 a menos de 34
34 a menos de 42
42 a menos de 50
50 a menos de 58
58 a menos de 66
66 a menos de 74
74 a menos de 82
82 a menos de 90

Además, la especificación "a menos de" evita el traslape entre los límites superiores de cada clase y los límites inferiores de la clase siguiente. Falta determinar ahora a qué clase pertenece cada dato para establecer la frecuencia de cada clase. Al hacer esto se obtiene la tabla de clases y frecuencias correspondiente que se muestra en la página siguiente.

Edades x	Frecuencias f
18 a menos de 26	5
26 a menos de 34	8
34 a menos de 42	13
42 a menos de 50	21
50 a menos de 58	17
58 a menos de 66	8
66 a menos de 74	5
74 a menos de 82	2
82 a menos de 90	1
Suman las frecuencias, $\sum f$	80

El resumen del procedimiento anterior sería el siguiente:

- Determinar el número de clases o intervalos utilizando la raíz cuadrada del número de elementos de la serie.
- Establecer el rango de los datos: máximo – mínimo.
- Dividir el rango entre el número de clases para determinar los intervalos de clase (este paso normalmente requiere de redondeos y ajustes para llegar, de preferencia, a valores enteros y gruesos, como decenas, por ejemplo).
- Determinar cuántos de los datos caen dentro de cada clase.

Se presenta ahora un ejemplo de conversión de una serie simple en una de datos y frecuencias para ilustrar que, en este caso, no es conveniente convertir la serie en una de clases y frecuencias dado que la cantidad de datos y frecuencias es suficiente para hacer una compactación adecuada de los datos.

Ejemplo 2.3

Se examinaron 50 cajas de focos, cada una de las cuales contiene 100 piezas. En la tabla 2.7 se muestra el número de focos defectuosos encontrados en cada una de las 50 cajas.

Tabla 2.7 Número de focos defectuosos encontrados en 50 cajas de 100 focos cada una

x	x	x	x	x	x	x	x	x	x
0	2	1	0	3	0	2	0	0	1
4	2	5	1	0	1	1	0	0	0
1	0	1	1	0	0	0	2	0	1
2	1	0	0	1	1	0	0	0	1
1	0	1	2	3	0	1	0	1	1

Esta tabla mejoraría si se ordenaran las observaciones, como se hizo en el ejemplo 2.1. Los datos ordenados aparecen en la tabla 2.8 en la cual se aprecia mejor la cantidad de mayor frecuencia de focos defectuosos en las cajas.

Tabla 2.8 Datos ordenados de focos defectuosos encontrados en 50 cajas

x	x	x	x	x	x	x	x	x	x
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1
2	2	2	2	2	2	3	3	4	5

Ahora, en la tabla 2.9 se presentan estos mismos datos de focos defectuosos en una tabla de datos y frecuencias.

Tabla 2.9 Tabla de datos y frecuencias de focos defectuosos

Número de focos defectuosos x	Frecuencia f
0	22
1	18
2	6
3	2
4	1
5	1
Total	50

De nueva cuenta se aprecia que esta forma de presentación de datos es más compacta, al mismo tiempo que permite detectar cuál es el valor que más se repite (22 de las cajas no contienen ningún foco defectuoso). Al mismo tiempo se observa que no es necesario compactar aún más estos datos de focos defectuosos, por lo que no se construye una tabla de clases y frecuencias.

Así, las tablas de frecuencias son utilizadas cuando se tienen conjuntos numerosos de datos que pueden resumirse agrupando las observaciones que se repiten; las tablas simples se utilizan cuando se tienen conjuntos de datos poco numerosos (o cuando, teniendo muchos datos, resulta conveniente por alguna razón listarlos todos).

En la sección siguiente se presenta un ejemplo de una serie de frecuencias para datos cualitativos y un conjunto de ejercicios para que el estudiante practique lo revisado en esta sección.

2.2.2 Tablas de frecuencias para datos cualitativos

En la sección anterior se estudiaron tablas para series de valores cuantitativos agrupados en clases. Las tablas para series de valores cualitativos agrupados tienen la misma forma pero las clases no están dadas por intervalos numéricos sino por categorías. En el ejemplo siguiente se presentan datos dados en escala nominal.

EJEMPLO 2.4

La empresa Dulce Hogar, S.A. de C.V., se dedica a la fabricación y venta de vajillas. Al final del bimestre encontró que de las 30 000 que debió producir sólo 29 000 salieron completas. Las 1 000 restantes presentaron los defectos que a continuación se listan:

Defecto x	Cantidad f
Tazas despostilladas	300
Platos despostillados	137
Tazas rotas	120
Platos sin barniz	93
Falta de cajas	95

Defecto x	Cantidad f
Platos mal pintados	89
Tazas sin barniz	77
Platos rotos	49
Tazas mal pintadas	40
Total	1 000

En la tabla anterior se observa que la variable "defecto" es nominal (categorías sin orden), tiene 9 valores posibles (tazas despostilladas, platos despostillados, etc.) y cada valor posible tiene su correspondiente frecuencia asociada, que son los casos que caen dentro de cada clase.

En seguida se presenta otro ejemplo de distribución de frecuencias para datos cualitativos en escala nominal:

EJEMPLO 2.5

En la tabla siguiente se muestra la distribución de frecuencias de las acciones que se negociaron en la Bolsa Mexicana de Valores un día determinado, de acuerdo al sector de la economía al que pertenecen.

Sector (x)	Número de acciones Frecuencia (f)
Industria extractiva	4
Industria de transformación	37
Industria de la construcción	21

Sector (x)	Número de acciones Frecuencia (f)
Comercio	21
Comunicaciones y transporte	28
Varios	15
Servicios financieros	34
Total	160

Fuente: Periódico *El Financiero*, 1 de abril de 2011, pp. 11A-14AA.

2.2.3 Frecuencias absolutas, relativas y acumuladas

En los ejemplos anteriores de tablas de frecuencias, y de clases y frecuencias, se manejaron frecuencias absolutas, ya que simplemente se contaron los elementos que correspondían a cada valor o a cada clase. Sin embargo, en algunas aplicaciones conviene manejar las frecuencias de las series no en forma absoluta sino en forma relativa y también en forma relativa acumulada.

En las tablas de frecuencias acumuladas simplemente se suman las frecuencias de todas las categorías anteriores, de manera que la frecuencia de cualquiera de esas categorías es igual a la suma de las frecuencias absolutas de todas las categorías precedentes. A manera de ilustración se reproduce la tabla 2.9 en la tabla 2.10, que tiene los datos de focos defectuosos, incluyendo una columna de frecuencias relativas.

En la tabla 2.10 se incluyen 2 columnas de frecuencias relativas, las cuales se obtienen dividiendo la frecuencia de cada categoría (número de focos defectuosos) entre el total de los focos defectuosos. La

columna final de la tabla, de frecuencias relativas porcentuales, se obtiene multiplicando las frecuencias relativas por 100.

Tabla 2.10 Tabla de frecuencias de focos defectuosos, incluyendo frecuencias acumuladas y frecuencias relativas

Número de focos defectuosos *x*	Frecuencia *f*	Frecuencia acumulada	Frecuencia relativa (tanto por uno)	Frecuencia relativa (tanto por ciento)
0	22	22	0.44	44
1	18	40	0.36	36
2	6	46	0.12	12
3	2	48	0.04	4
4	1	49	0.02	2
5	1	50	0.02	2
Total	50		1.00	100

Una utilidad práctica fácilmente identificable de las frecuencias relativas y de las acumuladas es que puede saberse, por ejemplo, la proporción de cada número de artículos defectuosos: puede conocerse que 44% de las cajas no contiene focos defectuosos y que 36% tienen uno.

2.2.4 Tablas de doble entrada o de clasificación cruzada o de contingencias

Una tabla tiene una sola entrada cuando los datos que contiene corresponden sólo a una variable; es decir, cuando sus datos son mediciones de acuerdo con una característica. En los ejemplos que se presentaron hasta aquí —los de número de focos defectuosos y defectos en vajillas— se tenían datos con una sola variable.

Por otro lado, la tabla completa 2.1 de los salarios mínimos, que contiene los sueldos para las 3 áreas geográficas, es una tabla de 3 entradas o 3 variables, una por cada área. Esta tabla se convirtió en 3 tablas de una entrada cuando se manejaron separadamente los datos para cada área. Además, como los datos de cada área son independientes entre sí, la tabla de 3 entradas es una **tabla de datos no cruzados**.

Tabla de datos no cruzados. Tabla cuyos datos de cada área son independientes entre sí.

Se presenta en seguida una tabla de doble entrada, con datos cruzados. Esta tabla contiene datos clasificados de acuerdo con 2 variables y son de gran importancia en estadística.

El ejemplo se refiere a la población mexicana, de acuerdo al Censo de Población y Vivienda que el Instituto Nacional de Estadística, Geografía e Informática (INEGI) realizó en el año 2010, clasificada de acuerdo con 2 variables: edad y género. Los datos se presentan en la tabla 2.11.

Tabla 2.11 Edad y género de la población mexicana

Edad (años)	Hombres	Mujeres	Total
0 a 10	12 146 191	11 769 700	23 915 891
11 a 20	10 927 296	10 856 730	21 784 026
21 a 30	8 938 941	9 617 150	18 556 091
31 a 40	7 898 681	8 674 176	16 572 857
41 a 50	5 943 904	6 520 204	12 464 108
51 a 60	4 043 049	4 432 217	8 475 266
61 a 70	2 405 484	2 695 674	5 101 158
71 a 80	1 313 064	1 516 824	2 829 888
81 a 90	465 190	593 386	1 058 576
91 a 99	65 984	96 812	162 796
100 o más	7 228	11 247	18 475
Total	54 155 012	56 784 120	110 939 132

Fuente: INEGI, *XII Censo General de Población y Vivienda, 2010*, disponible en: http://www3.inegi.org.mx/sistemas/TabuladosBasicos/Default.aspx?c=27302&s=est, consultado el 23 de marzo de 2011.

Los 110 939 132 mexicanos que se contabilizan están clasificados de acuerdo con su género y con cada una de 11 categorías de edad.

2.2.5 Uso de Excel. Construcción de distribuciones de frecuencias

Excel Con este paquete pueden elaborarse distribuciones de frecuencia con facilidad, tanto series de datos y frecuencias como series de clases y frecuencias. En los ejemplos siguientes se ilustra el procedimiento.

Ejemplo 2.6

Elaboración de una serie de datos y frecuencias con Excel

Se presentan en la tabla siguiente los tiempos de duración (en horas) para muestra aleatoria de 50 baterías.

163	159	150	136	136	138	155	158	135	166
132	144	125	157	146	145	145	150	144	142
154	139	139	168	158	151	151	153	148	156
152	146	134	158	154	154	154	151	150	154
148	144	156	167	156	141	141	138	148	160

El primer paso consiste en copiar los datos a una columna de una hoja de trabajo de Excel, por ejemplo, la columna A. En seguida se ordenan de menor a mayor, con lo que pueden identificarse los valores que se repiten. Después se anotan en otra columna, por ejemplo la C, en orden ascendente, vaciando los valores una única vez. Los valores se repiten y se tienen los siguientes datos en las celdas C1 a C28:

125	151
132	152
134	153
135	154
136	155
138	156
139	157
141	158
142	159
144	160
145	163
146	166
148	167
150	168

En seguida se marcan las celdas D1 a D28, Excel las sombrea y deja en blanco la celda D1 que a la vez, contiene el cursor. En esta celda se anota la función FRECUENCIA, con la siguiente sintaxis:

=FRECUENCIA(DATOS,GRUPOS)

Como los datos están en las celdas A1 a A50 y GRUPOS corresponde a todos los valores diferentes que se anotan antes y que se encuentran en las celdas C1 a C28, la función que se anota en esa celda es:

=FRECUENCIA(A1:A50,C1:C28)

Es importante tener presente que después de teclear toda la función no simplemente se oprime la tecla "Enter" sino que se deben oprimir simultáneamente las teclas "Ctrl." y mayúsculas, a la vez que se oprime "Enter".

Una vez que se hace esto se obtiene la serie de datos y frecuencias en las celdas C1 a D28, que se muestra en la tabla siguiente en la cual se ha anotado la suma de las frecuencias en el último renglón:

Horas de duración de baterías	Frecuencia *f*
125	1
132	1
134	1
135	1
136	2
138	2
139	2
141	2
142	1
144	3
145	2
146	2
148	3
150	3
151	3
152	1
153	1
154	5
155	1
156	3
157	1
158	3
159	1
160	1
163	1
166	1
167	1
168	1
Suma	50

EJEMPLO 2.7

Elaboración de una serie de clases y frecuencias con Excel

Utilizando los mismos datos de horas de duración de baterías se tendrán los cincuenta datos en las celdas A1 a A50 y, entonces, se seguirá el procedimiento sugerido en el ejemplo 2.2:

1. Determine el número de clases o intervalos, utilizando la raíz cuadrada del número de elementos de la serie. Agrupe los datos en 7 clases.
2. Determine el rango de los datos: el primero y el último de los valores de la serie ordenada en Excel (máximo – mínimo) 168 – 125 = 43.
3. Divida el rango entre el número de clases para determinar los intervalos de clase, por lo que se utiliza un intervalo de amplitud 7. Entonces las clases serían:

120 a menos de 127
127 a menos de 134
134 a menos de 141
141 a menos de 148
148 a menos de 155
155 a menos de 162
162 a menos de 169

Para el paso final que determinará las frecuencias correspondientes a cada clase, en la hoja de Excel se anotan, por ejemplo, en las celdas F1 a F7 los límites superiores de las clases anteriores, sea:

127
134
141
148
155
162
169

Lo que hace el paquete para determinar las frecuencias correspondientes a cada clase es contar todos los valores iguales o menores a este límite superior. Una vez anotados esos valores se marcan las celdas contiguas, G1 a G7, como sucedió en el ejemplo anterior, Excel las sombrea y deja en blanco la celda G1 que, a la vez, contiene el cursor. Aquí se anota la función:

=FRECUENCIA(A1:A50,F1:F7)

De nueva cuenta se oprimen simultáneamente las teclas "Ctrl." y mayúsculas, a la vez que se oprime la tecla "Enter". Haciendo esto se obtiene la siguiente distribución de clases y frecuencias en las celdas F1 a G7:

127	1
134	2
141	9
148	11
155	14
162	9
169	4
Suma	50

NOTA

En la sección 1.9 del capítulo 1, donde se explica cómo se utilizará Excel en este texto, se menciona que en el complemento de Excel que se llama "Análisis de Datos", existe un procedimiento que se denomina "Histograma". Este último no se analiza aquí, que es donde correspondería, porque hace exactamente lo mismo que la función =Frecuencia que se acaba de aplicar.

Series simples, de datos y frecuencias, y de clases y frecuencias

EJERCICIOS 2.1

1. Elabore una tabla de datos y frecuencias con los siguientes datos de número de hijos por cada familia en una unidad habitacional con 70 familias. Una vez construida esa tabla, elabore una tabla de frecuencias relativas y otra de frecuencias acumuladas.

x	*x*	*x*	*x*	*x*	*x*	*x*
1	12	0	4	3	0	12
9	5	2	3	0	3	2
3	6	1	3	1	5	9

x	*x*	*x*	*x*	*x*	*x*	*x*
2	9	3	5	7	10	0
12	2	7	10	2	6	7
0	5	4	5	3	9	1
4	10	9	2	7	0	0
2	2	3	5	1	2	8
4	1	2	3	0	10	2
0	3	4	1	2	7	11

2. Elabore una tabla de clases y frecuencias con los siguientes datos de peso (en kg) de 65 pacientes. Una vez construida esa tabla realice una tabla de frecuencias relativas y otra de frecuencias acumuladas.

x	*x*	*x*	*x*	*x*	*x*	*x*
67	54	44	55	67	42	38
55	73	67	56	34	45	36
80	25	56	54	67	48	88
37	82	34	45	78	66	55
45	34	48	73	55	76	60
62	63	56	78	58	62	
65	85	53	65	87	33	
56	56	44	87	54	45	
40	67	72	33	77	40	
55	39	67	35	48	59	

3. Elabore una tabla de clases y frecuencias con los siguientes datos de promedio de calificaciones de un grupo de 50 alumnos de licenciatura. Una vez construida esa tabla realice una tabla de frecuencias relativas y otra de frecuencias acumuladas.

x	*x*	*x*	*x*	*x*
8.0	9.5	8.0	8.7	7.0
6.9	7.0	8.8	8.3	7.9
7.5	8.9	9.3	7.8	7.8
8.0	9.0	8.7	7.0	8.9
9.7	9.4	8.8	8.3	9.5
9.3	6.8	7.8	7.8	7.5
8.7	7.0	6.7	7.5	6.8
8.8	8.3	7.9	6.6	9.5
7.8	7.8	5.9	8.2	7.4
8.3	6.5	8.1	7.8	6.8

4. Elabore lo siguiente:

a) Una tabla de datos y frecuencias.
b) Una tabla de clases y frecuencias con los siguientes datos de rendimiento mensual de 100 sociedades de inversión en instrumentos de deuda. Una vez construida esa tabla elabore una tabla de frecuencias relativas y otra de frecuencias acumuladas.

x	*x*	*x*	*x*	*x*	*x*	*x*	*x*	*x*	*x*
0.38	0.60	0.35	0.38	0.33	0.41	0.37	0.38	0.35	0.36
0.41	0.41	0.39	0.41	0.37	0.43	0.39	0.33	0.36	0.38
0.35	0.42	0.22	0.35	0.37	0.35	0.37	0.30	0.36	0.36
0.39	0.39	0.33	0.35	0.40	0.39	0.40	0.60	0.38	0.35
0.26	0.27	0.24	0.32	0.36	0.36	0.37	0.31	0.36	0.34
0.34	0.31	0.36	0.35	0.38	0.35	0.31	0.14	0.35	0.35
0.37	0.33	0.35	0.31	0.25	0.34	0.38	0.34	0.34	0.36
0.34	0.36	0.29	0.36	0.31	0.39	0.37	0.41	0.34	0.38
0.28	0.35	0.35	0.30	0.37	.03	0.33	0.38	0.33	0.39
0.34	0.36	0.34	0.38	0.28	0.41	0.35	0.25	0.24	0.27

5. Elabore lo siguiente:

a) Una tabla de datos y frecuencias.
b) Una tabla de clases y frecuencias con los siguientes datos de tiempo de acceso a 122 modelos de disco duro para computadora, en milisegundos. Una vez construida esa tabla elabore una tabla de frecuencias relativas y otra de frecuencias acumuladas.

x	*x*	*x*	*x*	*x*	*x*	*x*	*x*	*x*	*x*	*x*
16	17	15	16	12	16	16	29	12	16	7
17	17	16	16	17	16	16	24	12	15	16
14	17	19	15	17	12	17	25	12	15	24
12	24	19	16	16	14	17	24	12	15	16
17	24	19	17	24	14	17	20	12	15	4
11	24	15	14	16	25	12	20	12	15	23
24	24	24	12	16	21	17	8	15	12	24
20	16	16	15	12	24	11	17	15	12	22
16	16	16	16	24	24	16	17	15	12	22
15	16	16	16	18	24	12	12	16	12	13
18	15	21	17	16	17	22	12	15	12	16

6. Elabore una tabla de clases y frecuencias con los siguientes datos de lapsos, en minutos, necesarios para 50 clientes de un banco comercial que realizan una transacción bancaria. Una vez construida esa tabla elabore una tabla de frecuencias relativas y otra de frecuencias acumuladas.

x	*x*	*x*	*x*	*x*
2.3	0.2	2.9	0.4	2.8
2.4	4.4	5.8	2.8	3.3
3.3	9.7	2.5	5.6	9.5
1.8	4.7	0.7	6.2	1.2
7.8	0.8	0.9	0.4	1.3
3.1	3.7	7.2	1.6	1.9
2.3	4.6	3.8	1.5	2.7
0.4	1.3	1.1	5.5	3.4
4.2	1.2	0.5	6.8	5.2
6.2	7.6	1.4	0.5	1.4

7. Elabore 4 tablas de clases y frecuencias con los siguientes datos de contenido de nutrientes por cada 100 gramos de alimentos: calorías, proteínas, grasas y carbohidratos. Una vez construida esa tabla elabore una tabla de frecuencias relativas y otra de frecuencias acumuladas.

	Energía (calorías)	Proteínas (gramos)	Grasas (gramos)	Carbohidratos (gramos)
Aceite	351.3	0.51	0.2	86.9
Aceituna verde	113.5	1.19	10.67	3.17
Aguacate	198.2	1.28	19.76	3.81
Apio	13.9	0.81	0.13	2.33
Arroz	350.5	7.59	0.31	79.31
Atún	294.1	23.87	22.07	0
Avena	395.8	14.19	7.39	68.11
Azúcar granulada	397.5	0	0	99.37
Betabel	34.1	1.19	0.07	7.17
Bizcocho corriente	303.4	7.8	5.39	55.92
Cacahuates	552.2	24.78	40.69	21.73
Calabacita	18.3	0.57	0.09	3.78
Camarones	160.2	23.76	10.12	0
Camote	107.4	1.54	0.59	23.94
Carnitas de cerdo	398.2	10.47	39.6	0
Catsup	109.3	2	0.4	24.46
Cebada perla	356.4	8.18	0.99	78.72
Cebolla	45.8	1.32	0.2	9.68
Chabacano	53	0.95	0.09	12.12
Chícharo	353.3	24.46	0.99	61.62
Chocolate amargo	569.6	5.5	52.84	17.97
Chuleta de carnero	243.1	12.23	21.58	0
Chuleta de cerdo	235.4	13.28	20.24	0
Ciruela	53.2	0.66	0.2	12.23
Coctel de frutas	77.7	0.4	0.2	18.57
Coco rallado	578.4	3.59	39.05	53.13
Cocoa	328.9	9	18.79	30.95
Col	20.9	1.01	0.15	3.85
Coliflor	13.9	1.08	0.09	2.2
Consomé en cubitos	258.5	17.69	0	46.95
Corazón de res	125.4	16.48	6.3	0.7
Hojuelas de maíz	358.8	7.9	0.7	80.21
Costilla de ternera	145.4	15.09	9.46	0
Costillas de res	195.8	15.27	14.96	0
Crema fresca	207.5	2.9	20	4
Durazno	44.9	0.44	0.09	10.56
Ejote	37.8	2.16	0.18	6.93
Elote	40.9	1.41	0.46	7.81

	Energía (calorías)	Proteínas (gramos)	Grasas (gramos)	Carbohidratos (gramos)
Espárrago	19.8	1.65	0.15	2.93
Espinaca	20.2	1.89	0.24	2.62
Filete de pescado	82.5	15.93	2.09	0
Fresa	39.4	0.77	0.57	7.77
Frijol	349.4	21.98	4.69	60.83
Galletas saladas	421.1	9.48	10.3	72.62
Garbanzo	368.5	20.77	4.69	60.83
Gelatina en polvo	392	9.39	0	88.59
Haba verde	52.6	2.99	3.3	9.42
Harina de maíz	364.8	9.09	3.7	73.81
Harina de trigo	354.4	10.78	0.9	75.81
Helado de crema	209.7	4	12.28	20.77
Hígado de res	131.3	19.78	4.2	3.59
Huevo	139.9	11.37	10.23	0.62
Jalea	260.5	0.2	0	64.92
Jamón	291.7	13.05	26.62	0
Jamón serrano	333.1	14.7	30.36	0.26
Jitomate o tomate	20	0.88	0.26	3.52
Jocoque	35.4	3.5	0.11	5.1
Jugo de naranja	55	0.51	0.11	12.89
Jugo de piña	54.1	0.31	0.11	12.98
Jugo de toronja	41.4	0.51	0.2	9.39
Leche condensada	326.7	8.1	8.38	54.74
Leche entera fresca	68.6	3.5	3.9	4.88
Leche evaporada	138.9	7	7.9	9.88
Lechuga	12.5	0.84	0.13	2
Leche descremada	35.4	3.5	0.11	5.1
Lengua de res	192.5	15.55	14.3	0.37
Limón	40.5	0.62	0.07	9.33
Maíz	356.4	8.49	0.79	78.8
Manzana amarilla	56.8	0.26	0.35	13.11
Manzana roja	389.6	1.48	0.99	93.79
Melón	10.6	0.29	0.09	2.16
Mermelada	287.5	0.51	0.31	70.71

(*continúa*)

(*continuación*)

	Energía (calorías)	Proteínas (gramos)	Grasas (gramos)	Carbohidratos (gramos)
Miel de abeja	318.8	0.31	0	79.4
Molida de res	315.3	15.97	27.94	0
Naranja	36.1	0.64	0.15	8.05
Nueces	572.7	7.22	56	9.98
Ostiones	49.5	5.98	1.19	3.7
Pan blanco	260.9	8.49	2	52.23
Pan de centeno	262.7	6.4	3.39	51.63
Papa	71.5	1.67	0.09	16.02
Pasta para sopa	359.9	12.98	1.41	73.81
Pastel	326.9	6.4	8.18	56.94
Pastel de manzana	265.8	2.9	9.59	41.95
Paté de hígado	257.8	16.68	20.57	1.5
Pavo	175.3	13.49	13.49	0
Pepinillos encurtidos	11.4	0.51	0.2	1.89
Pepino	10.1	0.48	0.07	1.89
Peras	75.2	0.2	0.11	18.37
Pescado entero	47.1	9.11	1.19	0
Pierna de carnero	190.5	14.94	14.52	0
Pierna de cerdo	291.7	13.05	26.62	0
Piloncillo	239.8	0	0	59.93
Pimiento	24.6	1.01	0.18	4.77
Piña entera	30.8	0.22	0.11	7.26
Piña rebanada	86.9	0.1	0.11	21.08
Plátano	65.8	0.79	0.13	15.38
Pollo	118.4	12.32	7.68	0
Pulpa de res	235	18.17	18.04	0
Pulpa de ternera	159.1	19.47	9.02	0
Queso amarillo	392.5	23.87	32.25	1.69
Queso crema	366.7	7.08	36.85	1.69
Rábano	11	0.59	0.04	2.05
Requesón	101	19.18	0.79	4.29
Salchicha de puerco	445.9	10.78	44.75	0
Salchicha de ternera	209.4	15.97	16.17	0
Salchicha para hot dog	200.7	15.18	14.08	3.3
Salchichón	216.5	14.78	15.88	3.59
Salmón enlatado	168.5	20.57	9.59	0
Sandía	14.3	0.22	0.09	3.17
Sardina en aceite	330.9	21.03	26.97	0.99
Carne seca de res	193.6	34.25	6.3	0
Tocino	624.8	9.08	64.9	1.1
Toronja	29.3	0.33	0.13	6.67
Uva	71.3	0.77	0.4	16.17
Uva pasa	298.1	2.29	0.51	71.1
Zanahoria	39.4	1.06	0.26	8.18

2.3 Gráficas (con Excel)

En esta sección se revisan los principales tipos de gráficas: de barras e histogramas, de círculo o de pastel y de línea, y dada la facilidad y versatilidad que ofrecen los paquetes de computación para elaborarlas se describirán los procedimientos para elaborarlas en Excel.

2.3.1 Histogramas

Una forma común de presentación gráfica son los **histogramas**, que son gráficas de barras en las que se usa un plano cartesiano (un plano con un eje vertical, que corresponde al eje de las y y un eje horizontal, el del eje de las x). En este plano se utiliza el eje horizontal para medir la variable y el eje vertical para medir la frecuencia, sea absoluta o relativa. Un conjunto de datos que pueden representarse mediante un histograma son las distribuciones de frecuencias de los salarios mínimos por área geográfica que se elaboraron antes y que se presentaron en las tablas 2.4, 2.5 y 2.6. Se reproducen en seguida esas tablas:

Histogramas. Gráficas de barras en las que se usa un plano cartesiano (un plano con un eje vertical, el eje de las y y un eje horizontal, el eje de las x).

Tabla 2.12 Serie de clases y frecuencias con los salarios mínimos vigentes en 2011 en el área A

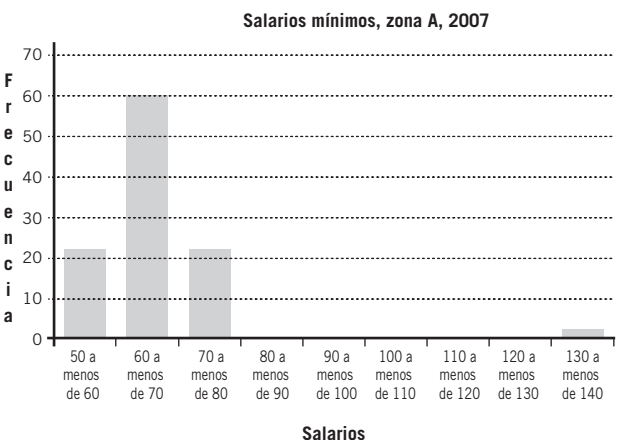
Clases de salarios *x*	Frecuencia *f*
70 a menos de 80	19
80 a menos de 90	44
90 a menos de 100	7
100 a menos de 110	0
110 a menos de 120	0
120 a menos de 130	0
130 a menos de 140	0
140 a menos de 150	0
150 a menos de 160	0
160 a menos de 170	0
170 a menos de 180	2
	72

Tabla 2.13 Distribución de frecuencias agrupadas de los salarios mínimos vigentes en 2011 en la zona B

Clases de salarios *x*	Frecuencia *f*
70 a menos de 80	29
80 a menos de 90	39
90 a menos de 100	2
100 a menos de 110	0
110 a menos de 120	0
120 a menos de 130	0
130 a menos de 140	0
140 a menos de 150	0
150 a menos de 160	0
160 a menos de 170	0
170 a menos de 180	2
	72

Tabla 2.14 Distribución de frecuencias agrupadas de los salarios mínimos vigentes en 2011 en la zona C

Clases de salarios *x*	Frecuencia *f*
70 a menos de 80	43
80 a menos de 90	26
90 a menos de 100	1
100 a menos de 110	0
110 a menos de 120	0
120 a menos de 130	0
130 a menos de 140	0
140 a menos de 150	0
150 a menos de 160	0
160 a menos de 170	2
170 a menos de 180	0
	72



Fuente: *Diario Oficial de la Federación*, disponible en: http://www.dof.gob.mx/nota_detalle.php?codigo=4941892/, consultado el 23 de marzo de 2011.

Figura 2.1 Distribución de frecuencias agrupadas de los salarios mínimos de la zona A.

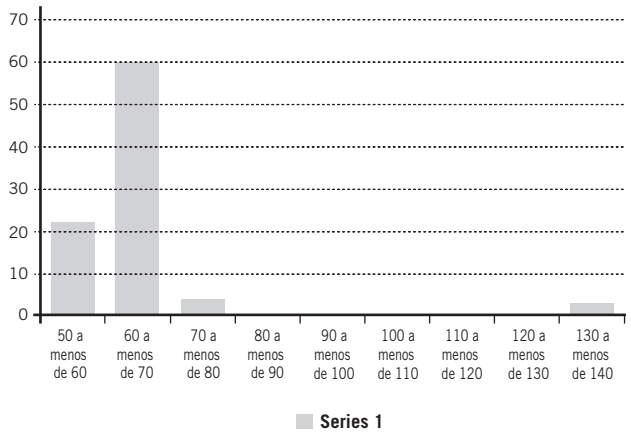
Figura 2.2 Cinta de opciones de "Insertar" de Excel 2007.

En la figura 2.1 se muestra la gráfica de los datos de la tabla 2.4 dibujada con Excel, y cuyos datos corresponden a los salarios mínimos para la zona A.

Para construir esta gráfica se copian y se marcan las 2 columnas de la tabla a una hoja de Excel. Después se da clic en la pestaña "Insertar" de la cinta de opciones y entonces se despliega la opción "Gráficos", tal como puede apreciarse en la figura 2.2.

Como puede verse en la figura 2.2, la sección de gráficos de la cinta de opciones de "Insertar" tiene como principal alternativa (el ícono más grande) una gráfica de "Columna", junto con "Línea", "Área", "Circular", "Dispersión", "Barra" (columna horizontal) y "Otros gráficos". Al acceder a cualquiera de estas opciones se muestran ilustraciones de los diversos tipos y conformaciones dentro de cada categoría, y además aparece la opción "Todos los tipos de gráfico…" que incluyen los mencionados y "Cotizaciones", "Superficie", "Anillos" y "Radial", todas éstas con ilustraciones que muestran el tipo de gráfica que cada alternativa produce.

Para construir el histograma de los salarios mínimos de cada una de las 3 zonas geográficas se elige la gráfica "Columna", la cual a su vez contiene 5 alternativas: "Columna en 2-D", "Columna en 3-D", "Cilíndrico",



Fuente: *Diario Oficial de la Federación*, disponible en: http://www.dof.gob.mx/nota_detalle.php?codigo=4941892/, consultado el 23 de marzo de 2011.

Figura 2.3 Primera versión de la gráfica de los salarios mínimos de la zona A.

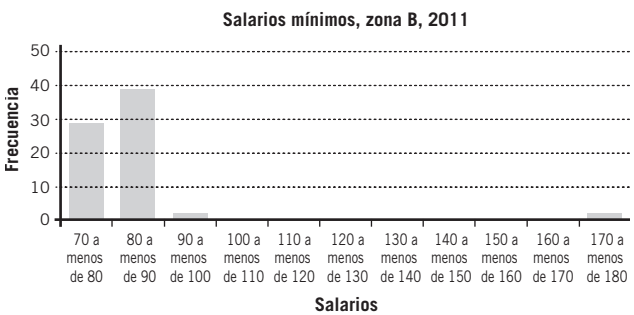


Figura 2.5 Distribución de frecuencias agrupadas de los salarios mínimos de la zona B.

Figura 2.4 Salarios mínimos, incluyendo título de la gráfica y leyendas de los ejes.

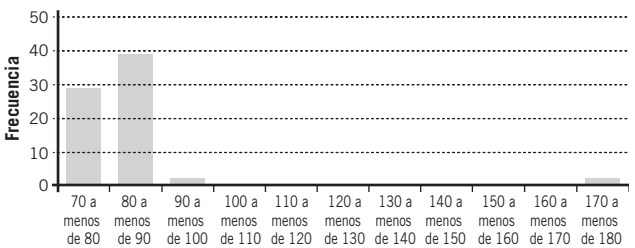


Figura 2.6 Distribución de frecuencias agrupadas de los salarios mínimos de la zona C.

"Cónico" y "Pirámide". Al elegir la primera opción, "Columna en 2-D", aparece la primera versión de la gráfica, misma que se reproduce en la figura 2.3.

Para mejorar dicha figura se elimina el texto del lado derecho y se le incluyen un título y leyendas a los ejes, eligiendo la pestaña "Presentación" de la cinta de opciones que aparece cuando se selecciona la gráfica. La cinta "Presentación" tiene las opciones de "Selección actual", "Etiquetas", "Ejes", "Fondo", "Análisis" y "Propiedades". En la sección "Etiquetas" se puede elegir la alternativa de "Rótulos de eje", desde donde se pueden escoger "Título de eje horizontal primario" y "Título de eje vertical primario". Con estas alternativas se pueden titular los ejes como "Salarios" y "Frecuencia" respectivamente, como aparecen en la figura 2.4. También en la sección "Etiquetas" se puede seleccionar la opción de "Título del gráfico" para anotar la leyenda que también se ve en la figura 2.7.

En la figura 2.5 se muestra la gráfica de los salarios mínimos de la zona B pero esta vez se han ensanchado las columnas hasta que ya no hay espacio entre ellas. Esto se hace dando clic al botón derecho sobre cualquiera de las barras de frecuencias con lo que aparecerá un menú que, hasta el final, incluye la opción "Dar formato a serie de datos…", y aquí puede elegirse cero como ancho del intervalo.

Finalmente, la figura 2.6 es la gráfica de los salarios mínimos para la zona C, en un formato muy simple, sin título y sin rótulos en los ejes.

EJEMPLO 2.8

La empresa Dulce Hogar, S.A. de C.V., se dedica a la fabricación y venta de vajillas. Al final del bimestre se encontró que de las 30 000 vajillas que debió producir sólo 29 000 salieron completas, mientras que las 1 000 restantes presentaron los defectos que se enlistan en la tabla de la página siguiente.

En la tabla se puede ver que la variable "defecto" es nominal (categorías sin orden), tiene 9 valores posibles (tazas despostilladas, platos despostillados, etc.) y cada valor posible tiene su frecuencia asociada correspondiente, son aquellos casos que caen dentro de cada clase. En la figura 2.7 se muestra una gráfica de barras de este conjunto de datos.

Defecto *x*	Cantidad *f*
Tazas despostilladas	300
Platos despostillados	137
Tazas rotas	120
Platos sin barniz	93
Falta de cajas	95
Platos mal pintados	89
Tazas sin barniz	77
Platos rotos	49
Tazas mal pintadas	40
Total	1 000

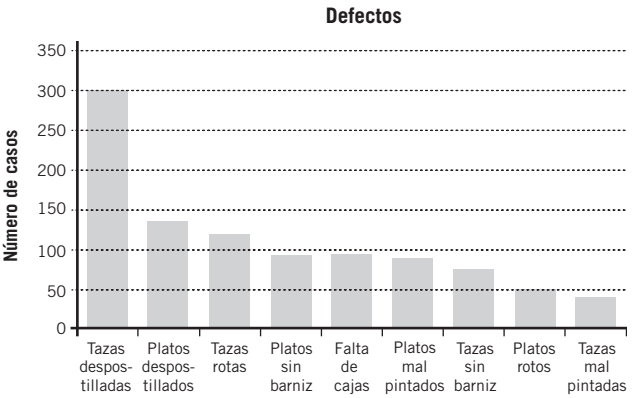


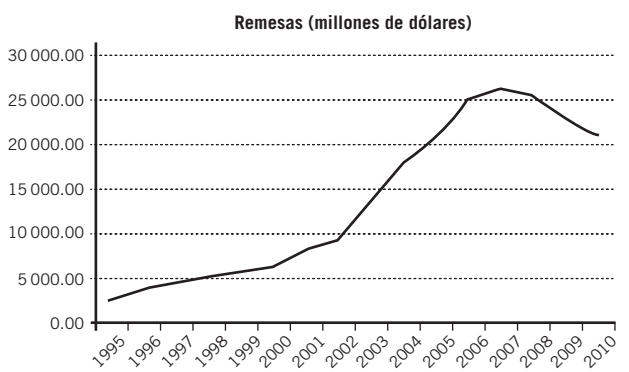
Figura 2.7 Gráfica de barras de los datos de defectos en vajillas.

2.3.2 Gráficas de líneas

En la figura 2.8 se muestra un ejemplo de gráfica de línea con datos anuales de las remesas internacionales que llegan al país. La gráfica corresponde a los siguientes datos:

Periodo	Remesas (millones de dólares)
1995	2 818.33
1996	4 058.31
1997	4 739.76
1998	5 242.06
1999	5 874.92
2000	6 327.52
2001	8 459.08
2002	9 551.01
2003	13 712.01
2004	17 666.32
2005	20 608.12
2006	25 030.16
2007	25 979.14
2008	25 257.57
2009	22 653.94
2010	20 964.00

Fuente: Banco de México, disponible en: http://www.banxico.gob.mx/polmoneinflacion/estadisticas/balanzaPagos/balanzaPagos.html, consultado el 24 de marzo de 2011.



Fuente: Banco de México, disponible en: http://www.banxico.gob.mx/polmoneinflacion/estadisticas/balanzaPagos/balanzaPagos.html, consultado el 24 de marzo de 2011.

Figura 2.8 Remesas internacionales anuales recibidas en México en los años 1995 a 2010.

El constante crecimiento de estas remesas y el inicio de su disminución en 2008 puede observarse en esta gráfica a partir de que el Banco de México mantiene estas estadísticas.

Se hizo esta gráfica en Excel, utilizando la opción "Línea". Los detalles restantes para conformar la gráfica son iguales a los que se describieron para los histogramas. Vale la pena mencionar que los datos y la gráfica corresponden a lo que en estadística se conoce como **series de tiempo**, en las que se tienen datos registrados en el tiempo, años en este caso, que se trazan sobre el eje horizontal, en tanto que los valores de las observaciones (las remesas) se miden sobre el eje vertical. Se analiza este tema con detalle en el capítulo 16.

Series de tiempo. Series de datos registrados en el tiempo que se trazan sobre el eje horizontal, en tanto que los valores de las observaciones se miden sobre el eje vertical.

2.3.3 Histogramas y polígonos de frecuencias

En la tabla 2.15 se muestra la distribución de frecuencias de los rendimientos diarios de días hábiles consecutivos obtenidos de los precios de las acciones Alfa A, de la empresa del mismo nombre, durante las

operaciones en la Bolsa Mexicana de Valores del 3 de enero de 1990 y hasta el 23 de marzo de 2011, para un total de 5 322 días hábiles y 5 321 rendimientos. Aunque la información se obtuvo de una base de datos armada por el autor de este texto, vale la pena anotar que también se encuentran en la sección de finanzas del sitio de internet de Yahoo!

Con los datos de cierre de cada día se calcularon los rendimientos diarios dividiendo el precio de cierre de cada día entre el precio del día anterior y después se resta 1 a este cociente para, finalmente, multiplicar la diferencia por 100 y obtener los rendimientos porcentuales diarios mostrados en la tabla. Todo esto, por supuesto, en Excel. Siguiendo ese procedimiento se elaboró el histograma de la figura 2.9.

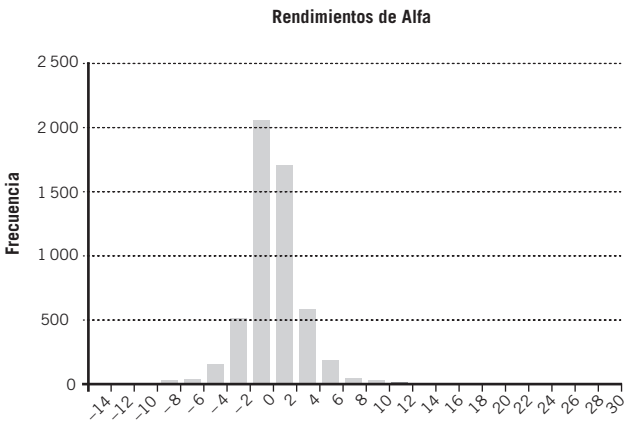


Figura 2.9 Histograma de los rendimientos de 5 321 días hábiles de Alfa A en la Bolsa Mexicana de Valores.

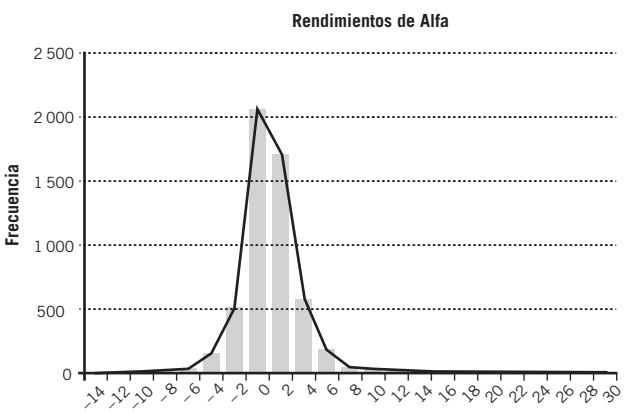


Figura 2.10 Histograma y polígono de frecuencias de los rendimientos teóricos diarios de 5 321 días hábiles consecutivos calculados sobre los precios de Alfa A en la Bolsa Mexicana de Valores.

Tabla 2.15 Rendimientos teóricos diarios de las acciones Alfa A en la Bolsa Mexicana de Valores, del 3 de enero de 1990 al 23 de marzo de 2011 (5 322 días hábiles y 5 321 rendimientos)

Rendimiento (%) *x*	Número de días *f*
–14	1
–12	2
–10	5
–8	24
–6	34
–4	154
–2	506
0	2 048
2	1 701
4	577
6	179
8	47
10	23
12	11
14	3
16	1
18	2
20	0
22	1
24	0
26	1
28	0
30	1
f	5 321

Las clases en esta distribución se dejaron tal como las maneja Excel, de manera que el 30 de la última clase dice realmente "Mayor de 28 y hasta 30".

Para trazar un polígono de frecuencias sobre el histograma de la figura 2.9, lo que se hace es marcar el centro de la parte superior de cada una de las barras para unir estos puntos con líneas. Haciendo esto se obtiene la gráfica que se muestra en la figura 2.10.

Con Excel lo que se requiere es duplicar la columna de las frecuencias para graficar tanto el histograma como el polígono de frecuencias, de manera que los datos que se usaron para construir esta figura se muestran en la tabla superior de la página siguiente.

Después de representar los datos en "Columna", Excel grafica 2 columnas iguales, una junto a la otra. Para convertir una de las columnas de datos en un polígono de frecuencias, se da clic derecho en alguna de las 2 barras, con lo que se tiene un menú que permite "Cambiar tipo de gráfica de series…" y aquí basta con marcar el tipo de gráfica de línea.

–14	1	1
–12	2	2
–10	5	5
–8	24	24
–6	34	34
–4	154	154
–2	506	506
0	2048	2048
2	1701	1701
4	577	577
6	179	179
8	47	47

10	23	23
12	11	11
14	3	3
16	1	1
18	2	2
20	0	0
22	1	1
24	0	0
26	1	1
28	0	0
30	1	1

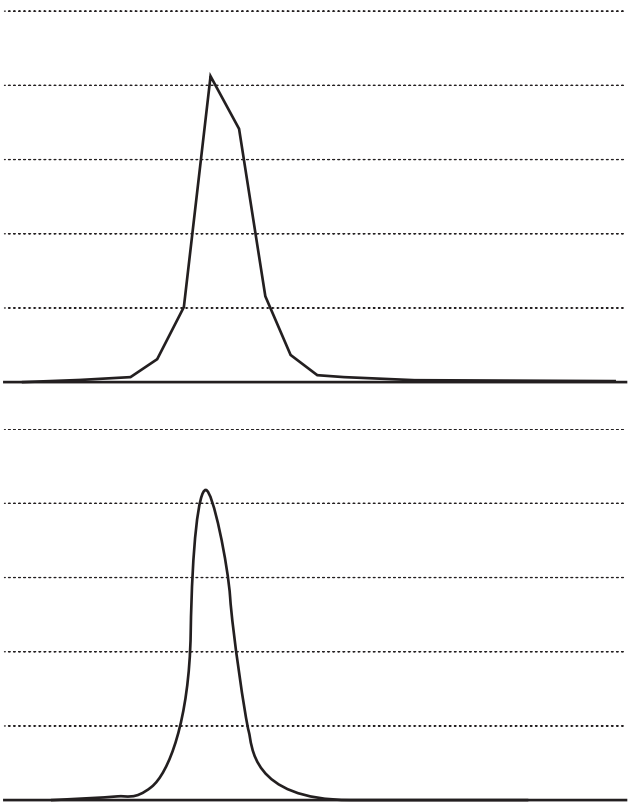


Figura 2.11 Gráfica de rendimientos de Alfa A, con un tipo de "líneas suavizadas".

Después de eliminar la leyenda innecesaria que aparece del lado derecho de la figura se tiene la gráfica que aparece en la figura 2.10. En esta figura se observa que el polígono de frecuencias se asemeja considerablemente a la forma de una campana, la cual corresponde a una distribución de probabilidad enormemente útil, conocida también como *campana de Gauss* o *distribución normal*, la cual se analiza con detenimiento en el capítulo 6, que se ocupa de las distribuciones continuas de probabilidad.

Finalmente y con el doble propósito de, por un lado, ilustrar más claramente la campana de Gauss, y por el otro, explicar un tipo más de gráfica de línea que se puede dibujar con Excel, en la figura 2.11 se muestran los mismos datos de rendimientos de Alfa A (después de eliminar una de las 2 series iguales de datos), pero ahora con una gráfica de "Dispersión" dando clic en el ícono en el que se ven los datos graficados como líneas suaves. En esta gráfica se aprecia con mayor claridad que esta serie (distribución) de datos se asemeja a una campana.

2.3.4 Gráficas circulares

Gráficas circulares o **de pastel.** Gráficas que son útiles para visualizar la composición de un conjunto de datos.

Las **gráficas circulares** o **de pastel** son útiles para visualizar la composición de un conjunto de datos. En seguida, algunos ejemplos.

EJEMPLO 2.9

Se reproducen como tabla 2.16 los datos de la tabla 2.11, que contiene la tabla de contingencias sobre el género y edad de la población mexicana, según el censo de 2010, y en la figura 2.12 se muestra la composición de la población por edades, en la que se aprecia gráficamente la superioridad numérica de las mujeres.

En la figura 2.12 (ver página siguiente) se aprecia gráficamente que las mexicanas son mayoría.

Tabla 2.16 Estado conyugal y género de la población mexicana de 12 años de edad o más

Edad (años)	Hombres	Mujeres	Total
0 a 10	12 146 191	11 769 700	23 915 891
11 a 20	10 927 296	10 856 730	21 784 026
21 a 30	8 938 941	9 617 150	18 556 091
31 a 40	7 898 681	8 674 176	16 572 857
41 a 50	5 943 904	6 520 204	12 464 108
51 a 60	4 043 049	4 432 217	8 475 266
61 a 70	2 405 484	2 695 674	5 101 158
71 a 80	1 313 064	1 516 824	2 829 888
81 a 90	465 190	593 386	1 058 576
91 a 99	65 984	96 812	162 796
100 o más	7 228	11 247	18 475
Total	54 155 012	56 784 120	110 939 132

Fuente: INEGI, *XII Censo General de Población y Vivienda, 2000*, disponible en: http://www3.inegi.org.mx/sistemas/TabuladosBasicos/Default.aspx?c=27302&s=est, consultado el 23 de marzo de 2011.

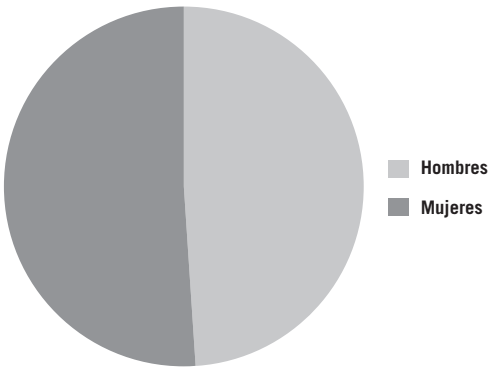


Figura 2.12 Composición de la población mexicana de acuerdo con su género.

El procedimiento en Excel, muy similar a los ya descritos para gráficas de barras: consiste en marcar en una hoja los datos que se desea graficar, incluyendo los nombres de las categorías y seleccionar "Circular" en el tipo de gráfico. Es posible, entre otras alternativas, hacer una gráfica en la que las "rebanadas" estén separadas.

2.3.5 Otras aplicaciones

Por supuesto son abundantes las aplicaciones de las gráficas. Sólo para ahondar un poco más sobre el punto, se aprovecha esta sección para mencionar las gráficas para control de calidad, entre las que se encuentran los diagramas de Pareto y las gráficas de control. Sin embargo, como más adelante se dedica todo un capítulo acerca del tema, se deja el tratamiento detallado para entonces.

Otro ejemplo muy común de gráficas son las que se construyen con datos de precios de los mercados financieros. Son particularmente populares las de acciones o de índices accionarios, en las que se grafican 3 precios diarios (en días hábiles): precio máximo, precio mínimo y precio de cierre o último. Excel permite graficar este tipo de datos.

Ejemplo 2.10

Gráficas de cotizaciones bursátiles

Como se mencionó, es fácil obtener datos de la sección de finanzas del sitio de internet de Yahoo! Esta sección se encuentra en http://mx.finance.yahoo.com/ en donde aparece un espacio para búsqueda de cotizaciones bursátiles. Por ejemplo, si se introduce la clave "^mxx", que corresponde al Índice de Precios y Cotizaciones de la Bolsa Mexicana de Valores, se llega a la página que contiene mucha información sobre este IPC; entre las opciones disponibles se encuentra la de acceder a datos históricos. De aquí se obtuvieron los datos de la tabla 2.17.

Nuevamente en la opción de gráficos de la cinta "Insertar", se eligen "Otros gráficos" y aparecen las "Cotizaciones", con 4 tipos. La primera es la más sencilla y la más común; la gráfica que se obtiene es la de la figura 2.13, a la que se le ha ajustado la escala del eje vertical, haciendo clic derecho sobre la escala y adecuando los valores mínimos para que la imagen abarque todo el espacio disponible.

Tabla 2.17 Cotizaciones (niveles) del IPC de la BMV, obtenidos en Yahoo!

	Precios		
Fecha	Máximo	Mínimo	Cierre
08 feb. 11	37 566.05	37 301.15	37 565.65
09 feb. 11	37 570.36	36 910.76	36 986.94
10 feb. 11	37 018.18	36 614.96	36 652.13
11 feb. 11	37 034.25	36 523.74	37 011.49
14 feb. 11	37 082.90	36 900.34	36 998.93
15 feb. 11	37 177.00	36 835.18	36 951.25
16 feb. 11	37 179.50	36 946.50	37 074.93
17 feb. 11	37 290.65	37 024.53	37 226.43
18 feb. 11	37 664.78	37 191.22	37 522.30
21 feb. 11	37 550.38	37 170.03	37 170.03
22 feb. 11	37 245.58	36 763.91	36 781.55
24 feb. 11	36 627.56	36 318.55	36 446.56
25 feb. 11	36 888.51	36 445.18	36 880.20
28 feb. 11	37 199.63	36 880.20	37 019.70
01 mar. 11	37 166.39	37 001.63	37 001.63
02 mar. 11	36 912.13	36 689.84	36 863.53
03 mar. 11	37 210.64	36 689.84	37 126.24
04 mar. 11	37 203.06	36 760.73	36 900.84
07 mar. 11	36 945.26	36 572.98	36 603.30
08 mar. 11	36 825.68	36 568.41	36 688.12
09 mar. 11	36 690.56	36 430.14	36 450.19
10 mar. 11	36 690.56	35 770.09	35 891.41
11 mar. 11	36 092.53	35 671.93	36 091.22
14 mar. 11	36 690.56	35 809.35	36 205.76
15 mar. 11	36 690.56	35 819.44	35 942.35
16 mar. 11	36 058.54	35 559.95	35 655.31
17 mar. 11	36 057.04	35 540.59	35 621.68
18 mar. 11	35 820.93	35 418.50	35 418.50
22 mar. 11	35 925.65	35 358.87	35 925.35
23 mar. 11	36 546.62	35 933.73	36 546.62

En esta figura cada barra representa los datos de un día; la parte más alta de la barra mide el máximo, la parte más baja es el mínimo y la pequeña muesca del lado derecho es el precio de cierre.

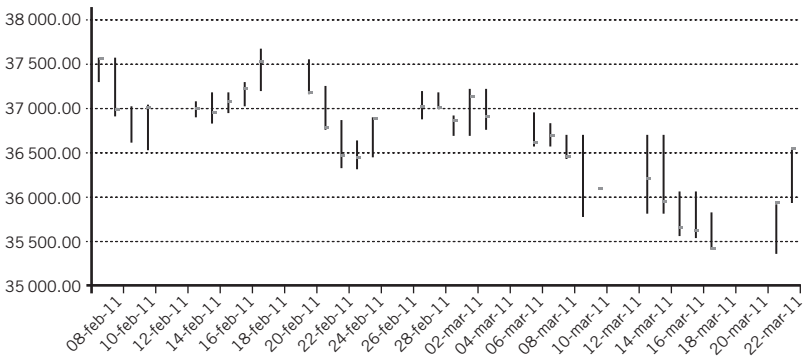


Figura 2.13 Gráfica de barras de cotizaciones del IPC de la BMV.

ejercicios 2.2 Gráficas

1. En el ejemplo 2.2 se construyó la siguiente tabla de clases y frecuencias con los datos de las edades de una muestra aleatoria de 80 empleados de una empresa.

Edades x	Frecuencias f
18 a menos de 26	5
26 a menos de 34	8
34 a menos de 42	13
42 a menos de 50	21
50 a menos de 58	17
58 a menos de 66	8
66 a menos de 74	5
74 a menos de 82	2
82 a menos de 90	1
Suman las frecuencias, Σf	80

Elabore, con Excel, un histograma y un polígono de frecuencias.

2. En el ejemplo 2.3 se construyó la siguiente tabla de clases y frecuencias con los datos de focos defectuosos encontrados en 50 cajas que contenían 100 focos cada una.

Número de focos defectuosos x	Frecuencia f
0	22
1	18
2	6
3	2
4	1
5	1
Total	50

Elabore un histograma y un polígono de frecuencias.

3. Con los datos del ejemplo 2.5, que se reproducen en seguida y que se refieren a la distribución de frecuencias de las acciones negociadas en la Bolsa Mexicana de Valores, por sector económico, en un día determinado, construya en Excel un histograma y un polígono de frecuencias.

Sector x	Número de acciones Frecuencia f
Industria extractiva	3
Industria de transformación	51
Industria de la construcción	20
Comercio	25
Comunicaciones y transportes	25
Varios	19
Servicios financieros	22
Total	165

4. Busque en internet datos apropiados para construir histogramas, elabórelos con Excel e incluya el polígono de frecuencias.

Gráficas de línea

5. En el cuadro siguiente aparecen los datos trimestrales del producto interno bruto de México, del primer trimestre de 2004 al cuarto trimestre de 2010. Elabore una gráfica de línea con estos datos.

Periodo	PIB (a precios de mercado)
1/2004	8 053 237 420
2/2004	8 500 417 248
3/2004	8 683 450 863
4/2004	9 062 187 350
1/2005	8 752 133 860
2/2005	9 200 787 839
3/2005	9 341 398 665
4/2005	9 712 629 613
1/2006	9 792 160 744
2/2006	10 407 069 212
3/2006	10 541 141 031
4/2006	10 775 992 918
1/2007	10 697 198 826
2/2007	11 212 225 694
3/2007	11 429 558 803
4/2007	11 944 362 185
1/2008	11 674 553 045
2/2008	12 433 567 241
3/2008	12 445 326 467
4/2008	12 247 081 422
1/2009	11 347 181 865
2/2009	11 596 797 823
3/2009	12 098 834 691
4/2009	12 675 254 863
1/2010	12 448 299 905
2/2010	12 942 872 938
3/2010	13 277 268 884
4/2010	13 880 244 589
Datos preliminares a partir de enero de 2009.	

Fuente: INEGI, disponible en: http://dgcnesyp.inegi.org.mx/cgi-win/bdieintsi.exe/Consultar, consultado el 23 de marzo de 2011.

6. En el cuadro siguiente se presentan los datos de la tasa unificada de desocupación que difunde el INEGI, con datos mensuales, de enero de 2001 a enero de 2011. Elabore con ellos una gráfica de línea.

Periodo	Total	Periodo	Total	Periodo	Total	Periodo	Total	Periodo	Total
Ene. 2001	2.71	Ene. 2003	3.08	Ene. 2005	4.2	Ene. 2007	3.96	Ene. 2009	5
Feb. 2001	3.23	Feb. 2003	3.01	Feb. 2005	3.81	Feb. 2007	4.02	Feb. 2009	5.3
Mar. 2001	2.7	Mar. 2003	3.14	Mar. 2005	3.6	Mar. 2007	4.01	Mar. 2009	4.76
Abr. 2001	2.63	Abr. 2003	3.01	Abr. 2005	3.71	Abr. 2007	3.6	Abr. 2009	5.25
Mayo 2001	2.54	Mayo 2003	2.87	Mayo 2005	3.29	Mayo 2007	3.23	Mayo 2009	5.31
Jun. 2001	2.48	Jun. 2003	3	Jun. 2005	3.65	Jun. 2007	3.26	Jun. 2009	5.17
Jul. 2001	2.97	Jul. 2003	3.83	Jul. 2005	4.12	Jul. 2007	3.95	Jul. 2009	6.12
Ago. 2001	2.58	Ago. 2003	4.09	Ago. 2005	3.68	Ago. 2007	3.92	Ago. 2009	6.28
Sep. 2001	2.8	Sep. 2003	4.19	Sep. 2005	3.69	Sep. 2007	3.87	Sep. 2009	6.41
Oct. 2001	2.99	Oct. 2003	3.94	Oct. 2005	3.57	Oct. 2007	3.93	Oct. 2009	5.94
Nov. 2001	2.81	Nov. 2003	3.77	Nov. 2005	3.01	Nov. 2007	3.46	Nov. 2009	5.26
Dic. 2001	2.64	Dic. 2003	2.94	Dic. 2005	2.83	Dic. 2007	3.4	Dic. 2009	4.8
Ene. 2002	3.58	Ene. 2004	4.01	Ene. 2006	3.53	Ene. 2008	4.04	Ene. 2010	5.87
Feb. 2002	3.06	Feb. 2004	4.12	Feb. 2006	3.72	Feb. 2008	3.91	Feb. 2010	5.43
Mar. 2002	3.05	Mar. 2004	3.94	Mar. 2006	3.43	Mar. 2008	3.8	Mar. 2010	4.81
Abr. 2002	2.97	Abr. 2004	3.88	Abr. 2006	3.31	Abr. 2008	3.61	Abr. 2010	5.42
Mayo 2002	2.84	Mayo 2004	3.45	Mayo 2006	2.88	Mayo 2008	3.24	Mayo 2010	5.13
Jun. 2002	2.77	Jun. 2004	3.65	Jun. 2006	3.33	Jun. 2008	3.55	Jun. 2010	5.05
Jul. 2002	3.08	Jul. 2004	4.11	Jul. 2006/	3.98	Jul. 2008	4.15	Jul. 2010	5.7
Ago. 2002	3.11	Ago. 2004	4.44	Ago. 2006	4.03	Ago. 2008	4.15	Ago. 2010	5.44
Sep. 2002	3.2	Sep. 2004	4.18	Sep. 2006	3.98	Sep. 2008	4.25	Sep. 2010	5.7

Periodo	Total	Periodo	Total	Periodo	Total	Periodo	Total	Periodo	Total
Oct. 2002	2.93	Oct. 2004	4.08	Oct. 2006	3.89	Oct. 2008	4.11	Oct. 2010	5.7
Nov. 2002	2.81	Nov. 2004	3.61	Nov. 2006	3.58	Nov. 2008	4.47	Nov. 2010	5.28
Dic. 2002	2.34	Dic. 2004	3.52	Dic. 2006	3.47	Dic. 2008	4.32	Dic. 2010	4.94
								Ene. 2011	5.43

Fuente: INEGI, disponible en: http://dgcnesyp.inegi.org.mx/cgi-win/bdieintsi.exe/Consultar/, consultado el 23 de marzo de 2011.

7. En el siguiente cuadro aparecen los datos mensuales del Índice Nacional de Precios al Consumidor (INPC), de enero de 2001 a febrero de 2011 que difunde el Banco de México en su página de internet. Elabore con estos datos una gráfica de línea.

Mes	INPC	Mes	INPC	Mes	INPC
Ene. 2001	64.659787943150	Mayo 2004	74.864322509016	Sep. 2007	85.295111472765
Feb. 2001	64.616994979760	Jun. 2004	74.984311751360	Oct. 2007	85.627495465924
Mar. 2001	65.026393744009	Jul. 2004	75.180845855199	Nov. 2007	86.231579237724
Abr. 2001	65.354409466737	Ago. 2004	75.644942177598	Dic. 2007	86.588098998021
Mayo 2001	65.504375883541	Sep. 2004	76.270403343149	Ene. 2008	86.989442325860
Jun. 2001	65.659309337784	Oct. 2004	76.798631846800	Feb. 2008	87.248039830912
Jul. 2001	65.488710598360	Nov. 2004	77.453745526263	Mar. 2008	87.880396929930
Ago. 2001	65.876712887810	Dic. 2004	77.613731182722	Abr. 2008	88.080379000503
Sep. 2001	66.489951356085	Ene. 2005	77.616489556109	Mayo 2008	87.985215118645
Oct. 2001	66.790457310724	Feb. 2005	77.875087061160	Jun. 2008	88.349320405757
Nov. 2001	67.042057015578	Mar. 2005	78.226090074683	Jul. 2008	88.841690055374
Dic. 2001	67.134902470813	Abr. 2005	78.504685786792	Ago. 2008	89.354747505396
Ene. 2002	67.754636301573	Mayo 2005	78.307462089606	Sep. 2008	89.963658430623
Feb. 2002	67.711079179108	Jun. 2005	78.232296414804	Oct. 2008	90.576706915932
Mar. 2002	68.057434733438	Jul. 2005	78.538475860785	Nov. 2008	91.606269782709
Abr. 2002	68.429198616676	Ago. 2005	78.632260555950	Dic. 2008	92.240695661768
Mayo 2002	68.567893678498	Sep. 2005	78.947404715439	Ene. 2009	92.454469599277
Jun. 2002	68.902213711874	Oct. 2005	79.141180445891	Feb. 2009	92.658589229931
Jul. 2002	69.100011723087	Nov. 2005	79.710784550351	Mar. 2009	93.191644887010
Ago. 2002	69.362746788219	Dic. 2005	80.200395826581	Abr. 2009	93.517822540048
Sep. 2002	69.779950763035	Ene. 2006	80.670698489101	Mayo 2009	93.245433168061
Oct. 2002	70.087509395709	Feb. 2006	80.794135698179	Jun. 2009	93.417141911415
Nov. 2002	70.654355126782	Mar. 2006	80.895505920159	Jul. 2009	93.671601856385
Dic. 2002	70.961913759456	Abr. 2006	81.014115975809	Ago. 2009	93.895719694096
Ene. 2003	71.248784591726	Mayo 2006	80.653458655431	Sep. 2009	94.366711949963
Feb. 2003	71.446697882259	Jun. 2006	80.723107583458	Oct. 2009	94.652203595540
Mar. 2003	71.897691931068	Jul. 2006	80.944467047782	Nov. 2009	95.143194058464
Abr. 2003	72.020439546799	Ago. 2006	81.357533462517	Dic. 2009	95.536951859488
Mayo 2003	71.788046588927	Sep. 2006	82.178839138560	Ene. 2010	96.575479439774
Jun. 2003	71.847351616752	Oct. 2006	82.538117272245	Feb. 2010	97.134050050685
Jul. 2003	71.951480212119	Nov. 2006	82.971181894037	Mar. 2010	97.823643397489
Ago. 2003	72.167322929668	Dic. 2006	83.451138863412	Abr. 2010	97.511947204733

(*continúa*)

(*continuación*)

Mes	INPC	Mes	INPC	Mes	INPC
Sep. 2003	72.596939584727	Ene. 2007	83.882134705164	Mayo 2010	96.897519532732
Oct. 2003	72.863122616593	Feb. 2007	84.116596443078	Jun. 2010	96.867177425472
Nov. 2003	73.467895981740	Mar. 2007	84.298649086634	Jul. 2010	97.077503396247
Dic. 2003	73.783729734576	Abr. 2007	84.248308772317	Ago. 2010	97.347134394847
Ene. 2004	74.242309310200	Mayo 2007	83.837311137622	Sep. 2010	97.857433471482
Feb. 2004	74.686407425541	Jun. 2007	83.937991766255	Oct. 2010	98.461517243282
Mar. 2004	74.939488183818	Jul. 2007	84.294511526553	Nov. 2010	99.250412032025
Abr. 2004	75.052581492694	Ago. 2007	84.637929013261	Dic. 2010	99.742092088296
				Ene. 2011	100.228000000000
				Feb. 2011	100.604000000000

Fuente: Banco de México, *Índices de Precios al Consumidor y UDIS*, disponible en: http://www.banxico.org.mx/politica-monetaria-e-inflacion/estadisticas/inflacion/indices-precios.html, consultado el 23 de marzo de 2011.

8. Se reproduce en seguida la tabla 2.1, que es de doble entrada y contiene datos de 11 categorías de edad y género de la población mexicana de acuerdo con el Censo de población y vivienda de 2010.

Tabla 2.1. Edad y género de la población mexicana

Edad (años)	Hombres	Mujeres	Total
0 a 10	12 146 191	11 769 700	23 915 891
11 a 20	10 927 296	10 856 730	21 784 026
21 a 30	8 938 941	9 617 150	18 556 091
31 a 40	7 898 681	8 674 176	16 572 857
41 a 50	5 943 904	6 520 204	12 464 108
51 a 60	4 043 049	4 432 217	8 475 266
61 a 70	2 405 484	2 695 674	5 101 158
71 a 80	1 313 064	1 516 824	2 829 888
81 a 90	465 190	593 386	1 058 576
91 a 99	65 984	96 812	162 796
100 o más	7 228	11 247	18 475
Total	54 155 012	56 784 120	110 939 132

Fuente: INEGI, *XII Censo general de población y vivienda, 2000*, disponible en: http://www3.inegi.org.mx/sistemas/TabuladosBasicos/Default.aspx?c=27302&s=est, consultado el 23 de marzo de 2011.

Elabore gráficas circulares (de pastel) con los siguientes datos:

a) Edad de los hombres.
b) Edad de las mujeres.
c) Edad de la población total.
d) Género.

9. En el cuadro siguiente aparecen los siguientes datos: fecha, apertura, máximo, mínimo, cierre y volumen para varios días de las acciones Cemex CPO que cotizan en la Bolsa Mexicana de Valores.

Elabore los 4 tipos de gráficas con cotizaciones que permite la versión 2007 de Excel

Apertura	Máximo	Mínimo	Cierre	Volumen
11.66	11.73	11.43	11.45	15 904 800
11.55	11.73	11.55	11.55	30 535 300
11.59	11.66	11.5	11.64	14 674 600
11.66	11.75	11.56	11.57	23 464 900
11.55	11.75	11.45	11.45	4 490 800
11.53	11.53	10.92	10.98	55 548 000
11.05	11.18	10.76	11	43 841 400
10.97	11.16	10.81	10.88	42 939 700
10.98	11.1	10.88	10.98	29 992 200
10.98	11.11	10.83	10.87	30 445 900
10.95	10.95	10.79	10.77	4 366 000
10.6	10.92	10.6	10.7	22 584 700
10.6	10.92	10.58	10.58	31 379 830
10.65	10.7	10.49	10.62	20 661 100
10.7	10.7	10.4	10.53	22 667 600
10.5	10.76	10.41	10.68	21 789 900
10.52	10.54	10.23	10.39	52 962 200
10.33	10.45	10.13	10.32	42 954 500
10.3	10.51	10.24	10.46	24 535 900
10.49	10.58	10.33	10.51	20 750 100
10.3	10.57	10.24	10.57	16 912 624
10.59	10.59	10.2	10.32	41 256 000
10.48	10.51	10.25	10.29	28 332 000
10.31	10.38	10.27	10.32	56 518 600
10.49	10.49	10.31	10.43	24 245 000

Fuente: CEMEX-CPO, *Historical Prices*, disponible en: http://finance.yahoo.com/q/hp?s=CEMEXCPO.MX+Historical+Prices, consultado el 23 de marzo de 2011.

a) Máximo, mínimo y cierre.
b) Apertura, máximo, mínimo y cierre, como velas japonesas.
c) Volumen, máximo, mínimo y cierre.
d) Volumen, apertura, máximo, mínimo y cierre, como velas japonesas.

2.4 Resumen

Se revisaron las 2 principales formas de presentación de los datos estadísticos: las tablas y las gráficas. Se hizo hincapié en la utilidad de Excel para elaborar estas figuras. La cantidad posible de ambas formas de presentación es enorme, por ello se limitó este repaso a las más útiles y comunes; además, se comenzó por los principales elementos de las tablas y las gráficas, entre los que destacan el título, el cuerpo o la gráfica misma y la fuente de los datos.

Se ilustraron 3 tipos de tablas que pueden construirse a partir de un mismo conjunto de datos: las series simples, las series de datos y frecuencias, y las series de clases y frecuencias. Se incluyeron las tablas de frecuencias para datos cualitativos, otro tipo importante, y con respecto a todas las anteriores, se revisaron los casos de frecuencias absolutas, relativas y acumuladas.

Asimismo se vieron casos de tablas de doble entrada o de clasificación cruzada, también conocidas como *tablas de contingencias*.

En la sección de gráficas, donde Excel es especialmente útil, se generaron histogramas, gráficas de líneas —que se suelen utilizar en el análisis de series de tiempo como las que se estudian en el capítulo 16—, polígonos de frecuencias, gráficas circulares o de pastel y un ejemplo de otro tipo de gráficas: las de los precios de cotización de instrumentos que se negocian en bolsas de valores, como las acciones.

2.5 Ejercicios adicionales

1. En la tabla siguiente se muestran los porcentajes de crecimiento estimado del producto interno bruto (el valor de los bienes y servicios producidos dentro de cada país), según estimaciones de la CIA (Central Intelligence Agency) para 2010; puede observarse que México ocupa el lugar 64 de entre 216 países considerados. Con esos datos:

a) Construya una tabla de clases y frecuencias, incluyendo frecuencias absolutas y frecuencias relativas.

b) Elabore una gráfica de barras o histograma con la tabla de clases y frecuencias.

País	PIB	País	PIB	País	PIB	País	PIB
Seychelles	20.09	Chile	5.3	Israel	3.4	Isla de San Martín	1.6
Katar	19.4	Egipto	5.3	Túnez	3.4	Belice	1.5
Singapur	14.7	Burkina Faso	5.2	Australia	3.3	Noruega	1.5
Paraguay	14.5	Sierra Leona	5.2	Togo	3.3	Madagascar	1.5
Turkmenistan	11	Yemen	5.2	Libia	3.3	Kiribati	1.5
Congo, República	10.5	Sudán	5.2	Finlandia	3.2	Groenlandia	1.5
Taiwán	10.5	Mali	5.2	Jordania	3.2	Cuba	1.5
China	10.3	Isla de Man	5.2	Luxemburgo	3.2	Dominica	1.4
Afganistán	8.9	Camboya	5	Kuwait	3.2	Macedonia	1.3
Perú	8.7	México	5	Albania	3.1	El Salvador	1.2
Uruguay	8.5	Gambia	5	Unión de Myanmar	3.1	Islas Caimán	1.1
India	8.3	Mauritania	5	Botsuana	3.1	Santa Lucía	1.1
Mozambique	8.3	Islas Turcos y Caicos	4.9	Samoa Americana	3	Italia	1.1
Uzbekistán	8.2	Bielorrusia	4.8	Benin	3	Brunei	1
Timor Oriental	8	República de Yibuti	4.8	Guernsey	3	Dinamarca	1
Laos	7.8	Mundo	4.7	Guinea	3	Eslovenia	1
Tailandia	7.6	Ghana	4.7	Irán	3	Portugal	1
Argentina	7.5	Armenia	4.7	Japón	3	Macau	1
Brasil	7.5	Bermudas	4.6	Tuvalu	3	Islas Cocos	1
Panamá	7.5	Cabo Verde	4.5	Sudáfrica	3	Ecuatorial	0.9
Filipinas	7.3	Kosovo	4.5	Canadá	3	Granada	0.8
Turquía	7.3	Colombia	4.4	Camerún	2.8	Hungría	0.8
Líbano	7.2	Ucrania	4.3	Islas Salomón	2.8	Bosnia y Herzegovina	0.7
Malasia	7.2	República Dominicana	4.2	Nicaragua	2.8	Chipre	0.6
Etiopía	7	Marruecos	4.2	Polinesia Francesa	2.7	Islas Feroe	0.5
Kazajastán	7	Argelia	4.1	Estados Unidos	2.7	San Vicente y las Granadinas	0.5
Zambia	7	Namibia	4.1	Suiza	2.7	Lituania	0.4

(*continúa*)

(*continuación*)

País	PIB	País	PIB	País	PIB	País	PIB
Ribera Occidental	7	Vanuatu	4.1	Paquistán	2.7	Bulgaria	0.3
Sri Lanka	6.9	Suecia	4.1	Andorra	2.6	Islas Cook	0.1
Bután	6.8	Costa Rica	4	Somalia	2.6	España	-0.2
Hong Kong	6.8	Siria	4	Emiratos Árabes Unidos	2.6	Islas Marshall	-0.3
Vietnam	6.8	Eslovaquia	4	Guyana	2.5	Bahamas	−0.5
Nigeria	6.8	Maldivas	4	Honduras	2.5	Tonga	−0.5
Malawi	6.5	Kenia	4	Aruba	2.4	Islas Vírgenes Británicas	−0.6
Moldavia	6.5	Eritrea	4	Estonia	2.4	Barbados	−0.7
Tanzania	6.4	Baréin	3.9	República Checa	2.3	Jamaica	−0.8
Niue	6.2	Senegal	3.9	Guatemala	2.2	Corea del Norte	−0.9
Nueva Guinea	6.2	Burundi	3.9	Bélgica	2.1	Montserrat	−1
Corea del Sur	6.1	Bolivia	3.8	Trinidad y Tobago	2.1	Croacia	−1.4
Mongolia	6.1	Rusia	3.8	Nueva Zelanda	2.1	San Cristóbal y Nieves	−1.5
Bangladesh	6	Arabia Saudita	3.8	Austria	2	Irlanda	−1.6
República Democrática del Congo	6	Polonia	3.8	Chad	2	Letonia	−1.8
Sao SantoTome y Príncipe	6	Azerbaiján	3.7	Malta	2	Montenegro	−1.8
Ruanda	6	República Centroafricana	3.7	Islas Vírgenes	2	Rumania	−1.9
Liberia	6	Ecuador	3.7	Suazilandia	2	Samoa	−2
Indonesia	6	Costa de Marfil	3.6	Fiyi	1.8	Venezuela	−2.8
Gibraltar	6	Mauricio	3.6	Unión Europea	1.8	Islandia	−3.4
Angola	5.9	Omán	3.6	Guinea-Bissau	1.8	Kirguistán	−3.5
Zimbabue	5.9	Alemania	3.6	Liechtenstein	1.8	Antigua Barbuda	−4.1
Uganda	5.8	Curazao	3.5	Comoras	1.7	Grecia	−4.8
Georgia	5.5	Surinam	3.5	Holanda	1.7	Haití	−5.1
Irak	5.5	Nigeria	3.5	Serbia	1.7	Puerto Rico	−5.8
Tayikistán	5.5	Nepal	3.5	Francia	1.6	Anguila	−8.5
Gabón	5.4	Lesoto	3.5	Reino Unido	1.6	San Marino	−13

Fuente: Central Intelligence Agency, *The World Factbook*, disponible en:https://www.cia.gov/library/publications/the-world-factbook/rankorder/2003rank.html, consultado el 25 de marzo de 2011.

2. En la tabla siguiente se presenta la lista de las 50 ciudades más caras del mundo, según datos de Mercer, para 2007. Con los datos:

a) Construya una serie de clases y frecuencias con los datos del índice de costo de la vida.
b) Construya con esta serie un histograma.

Posición en marzo de 2008	Ciudad	País	Índice de costo de vida
1	Moscú	Rusia	142.4
2	Tokio	Japón	127
3	Londres	Reino Unido	125
4	Oslo	Noruega	118.3
5	Seúl	Corea del Sur	117.7
6	Hong Kong	China	117.6
7	Copenhague	Dinamarca	117.2
8	Ginebra	Suiza	115.8
9	Zúrich	Suiza	112.7
10	Milán	Italia	111.3
11	Osaka	Japón	110
12	París	Francia	109.4
13	Singapur	Singapur	109.1
14	Tel Aviv	Israel	105
15	Sídney	Australia	104.1
16	Dublín	Irlanda	103.9
16	Roma	Italia	103.9
18	San Petersburgo	Rusia	103.1
19	Viena	Austria	102.3
20	Beijing	China	101.9
21	Helsinki	Finlandia	101.1
22	Nueva York	Estados Unidos	100

Posición en marzo de 2008	Ciudad	País	Índice de costo de vida
23	Estambul	Turquía	99.4
24	Shanghái	China	98.3
25	Ámsterdam	Holanda	97
25	Atenas	Grecia	97
25	São Paulo	Brasil	97
28	Madrid	España	96.7
29	Praga	República Checa	96
30	Lagos	Nigeria	95.9
31	Barcelona	España	95.2
31	Río de Janeiro	Brasil	95.2
31	Estocolmo	Suecia	95.2
34	Douala	Camerún	95.1
35	Varsovia	Polonia	95
36	Melbourne	Australia	94.2
37	Múnich	Alemania	93.1
38	Berlín	Alemania	93
39	Bruselas	Bélgica	92.9
40	Frankfurt	Alemania	92.5
41	Dakar	Senegal	92.2
42	Kiev	Ucrania	91.7
43	Luxemburgo	Luxemburgo	91.3
44	Almaty	Kazakstán	90.7
45	Bratislava	Eslovaquia	90.6
46	Düsseldorf	Alemania	90.4
46	Riga	Letonia	90.4
48	Mumbai	India	90.3
49	Zagreb	Croacia	90
50	Hamburgo	Alemania	89.9

Fuente: Mercer, disponible en: http://www.mercer.com/costoflivingpr/, consultado el 25 de marzo de 2011.

3. En el siguiente cuadro se presentan los datos sobre género y estado civil de la población mexicana mayor de 12 años, con información del Censo de Población y Vivienda 2010, publicado por el INEGI.

	Hombres	Mujeres	Total
Total	40 947 872	43 979 596	84 927 468
Solteros	15 460 577	14 392 540	29 853 117
Casados	17 067 461	17 353 462	34 420 923
En unión libre	6 045 370	6 185 310	12 230 680
Separados	970 996	2 211 430	3 182 426
Divorciados	433 354	813 202	1 246 556
Viudos	819 019	2 914 338	3 733 357
No especificado	151 095	109 314	260 409

Fuente: INEGI, *Censo de Población y Vivienda, 2010*, disponible en: http://www3.inegi.org.mx/sistemas/TabuladosBasicos/Default.aspx?c=27302&s=est, consultado el 25 de marzo de 2011.

a) Convierta a proporciones los datos de hombres, mujeres y total.

b) Elabore 2 diagramas circulares (de pastel):

- uno para la población masculina,
- otro para la población femenina, y
- la población total, en donde se muestren los correspondientes porcentajes.

4. En el siguiente cuadro se presenta información del Banco Interamericano de Desarrollo sobre el crecimiento del PIB anual, por país latinoamericano, para el año 2010. Elabore una gráfica de barras con esos datos.

País	Valor
Argentina	9.24
Bolivia	4.06
Brasil	2.3
Chile	6.35
Colombia	5.13
Costa Rica	4.13
Ecuador	3.93
El Salvador	2.75
Guatemala	3.25
Honduras	4.58
México	2.96
Nicaragua	3.98
Panamá	6.38
Paraguay	2.7
Perú	6.67
República Dominicana	4.5
Uruguay	6.51
Venezuela	9.33

Fuente: Governance indicators database, disponible en: http://www.iadb.org/datagob/home_esp.html, consultado el 25 de marzo de 2011.

5. Se presenta información del INEGI sobre la población de 14 años y más para 2010. Elabore una gráfica de barras.

	Población de 14 años y más		
Periodo	Total	Población económicamente activa (PEA)	Población no económicamente activa (PNEA)
Ene.	100	58.27	41.73
Feb.	100	57.96	42.04
Mar.	100	58.61	41.39
Abr.	100	59.03	40.97
Mayo	100	59.44	40.56
Jun.	100	59.11	40.89
Jul.	100	59.06	40.94
Ago.	100	59.27	40.73
Sept.	100	58.52	41.48

(*continúa*)

(*continuación*)

	Población de 14 años y más		
Periodo	Total	Población económicamente activa (PEA)	Población no económicamente activa (PNEA)
Oct.	100	58.13	41.87
Nov.	100	57.93	42.07
Dic.	100	57.07	42.93

Fuente: INEGI, *Desocupación. Distribución porcentual de la población de 14 años y más según condición de actividad y ocupación nacional*, http://dgcnesyp.inegi.org.mx/cgi-win/bdiecoy.exe/618?s=est&c=25615, consultado el 25 de marzo de 2011.

6. En el siguiente cuadro se presentan los datos sobre las unidades económicas que arrojó el Censo Económico 2009, publicado por el INEGI, para las diferentes actividades económicas.

a) Convierta los datos a proporciones.
b) Elabore una gráfica con los datos obtenidos.

Unidades económicamente existentes en 2009 según actividad económica

	Total	
Actividad económica	absoluto	%
Total	5 144 056	
Comercio	2 424 249	
Servicios privados no financieros	2 056 437	
Industrias manufactureras	581 044	
Servicios financieros y de seguros	20 049	
Pesca y agricultura	19 454	
Construcción	19 020	
Transportes, correos y almacenamientos	18 257	
Minería	2 957	
Electricidad, agua y gas	2 589	

Fuente: INEGI, *Resumen de los resultados de los censos económicos 2009*, disponible en: http://www.inegi.org.mx/est/contenidos/espanol/proyectos/censos/ce2009/pdf/RD09-resumen.pdf, consultado el 25 de marzo de 2011.

7. En el siguiente cuadro se presentan los datos sobre la ventas totales de automóviles que arrojaron las estadísticas de AMIA 2010 (Asociación Mexicana de la Industria Automotriz, A.C.) y un cuadro con las ventas que fueron de exportación.

Producción total								
Periodo	Chrysler	Ford Motor	General Motors	Honda	Nissan	Toyota	Volkswagen	Total
Ene.	20 870	26 826	45 806	4 584	38 023	4 354	24 595	165 058
Feb.	20 766	25 780	43 005	4 200	37 086	4 354	32 101	167 292
Mar.	25 156	29 763	49 558	4 830	38 508	5 188	37 088	190 091
Abr.	19 771	25 338	45 159	4 067	37 449	4 496	33 997	170 277
Mayo	20 250	31 264	51 412	4 822	39 107	4 387	27 496	178 738
Jun.	23 408	39 484	50 212	4 971	45 966	4 730	37 424	206 195
Jul..	16 268	24 786	43 119	4 442	45 419	3 697	42 352	180 083
Ago.	23 989	39 662	52 416	5 148	39 946	4 844	39 735	205 740
Sep.	24 653	39 943	47 098	4 649	45 293	4 763	31 019	197 418
Oct.	25 222	43 171	46 872	4 612	52 278	4 710	43 843	220 708
Nov.	19 555	39 971	45 497	4 809	48 304	4 943	44 481	207 560
Dic.	17 411	27 661	39 196	3 867	39 115	3 812	40 554	171 616
Acumulado 2010	257 319	393 649	559 350	55 001	50 6494	54 278	434 685	2 260 776

Producción para exportación								
Periodo	Chrysler	Ford Motor	General Motors	Honda	Nissan	Toyota	Volkswagen	Total
Ene.	18 914	26 359	36 469	3 867	26 444	4 354	20 304	136 711
Feb.	18 174	25 340	33 843	3 505	25 360	4 354	27 472	138 048
Mar.	23 479	29 366	39 829	4 115	25 349	5 188	31 637	158 963
Abr.	17 799	24 898	34 973	2 607	29 497	4 496	29 555	143 825
Mayo	18 182	30 338	41 610	3 832	27 361	4 387	23 686	149 396
Jun.	20 520	38 149	42 311	4 140	30 181	4 730	33 449	173 480
Jul.	15 093	23 567	36 169	3 148	32 052	3 697	36 165	149 891
Ago.	22 360	37 821	45 083	4 281	24 343	4 844	34 107	172 839
Sep.	22 489	38 494	39 529	4 649	28 975	4 763	24 780	163 679
Oct.	21 910	42 265	38 308	2 154	35 426	4 710	31 150	175 924
Nov.	17 218	39 042	39 071	2 896	33 101	4 943	34 649	170 920

Producción para exportación								
Periodo	Chrysler	Ford Motor	General Motors	Honda	Nissan	Toyota	Volkswagen	Total
Dic.	14 129	27 642	34082	1 781	26 584	3 812	34 077	142 108
Acumulado 2010	230 267	383 281	46 1277	40 975	344 673	54 278	361 031	1 875 784

Fuente: http://www.amia.com.mx/autoprodtotal.php, consultado el 28 de marzo de 2011.

a) Obtenga los porcentajes para cada una de las cifras de enero a diciembre y totales.
b) Grafique los totales de la producción por exportación anual de cada empresa.

8. En el siguiente cuadro se presentan datos de los salarios mínimos generales promedio de 1996 a 2011 que publicó la Conasami.

a) Construya una tabla de clases y frecuencias, incluyendo frecuencias relativas.
b) Elabore un histograma con la tabla de clases y frecuencias.

Salario mínimo general promedio de los Estados Unidos Mexicanos 1996-2011	
Periodo	Pesos diarios
1996	18.43
1996	20.66
1996	24.3
1997	24.3
1998	27.99
1998	31.91
1999	31.91
2000	35.12
2001	37.57
2002	39.74
2003	41.53
2004	43.297
2005	45.24
2006	47.05
2007	48.88
2008	50.84
2009	53.19
2010	55.77
2011	58.06

Fuente: CONASAMI, *Salario mínimo general promedio de los Estados Unidos Mexicanos 1964-2012*, disponible en: http://www.conasami.gob.mx/pdf/salario_minimo/sal_min_gral_prom.pdf, consultado el 28 de marzo de 2011.

9. En el siguiente cuadro se muestran datos de activos netos de las Siefores (Sociedades de Inversión Especializadas en Fondos para el Retiro) que publicó la Comisión Nacional del Sistema de Ahorro para el Retiro (Consar).

Descripción del concepto	Ene. 2010	Feb. 2010	Mar. 2010	Abr. 2010	Mayo 2010	Jun. 2010	Jul. 2010	Ago. 2010	Sep. 2010	Oct. 2010	Nov. 2010	Dic. 2010
Activos netos de las Siefores básicas	1 150 391.30	1 186 654.20	1 213 394.20	1 228 845.50	1 235 168.60	1 270 928.40	1 307 326.40	1 330 597.60	1 367 628.40	1 405 063.30	1 364 397.50	1 373 663.00
Activos netos de las SB1	119 810.40	121 209.00	121 630.90	121 497.30	122 126.60	124 145.40	125 618.30	127 147.40	151 257.90	152 774.80	147 737.00	146 384.90
Activos netos de las SB2	279 720.30	286 292.50	291 071.50	293 853.60	295 465.70	302 698.10	309 983.20	314 696.80	334 977.80	342 461.00	333 754.80	335 238.90
Activos netos de las SB3	346 018.70	356 252.30	364 356.60	369 119.80	370 970.90	381 933.10	393 153.00	399 834.80	421 802.40	433 117.10	420 446.20	423 681.60
Activos netos de las SB4	331 971.60	344 469.70	354 313.50	360 081.20	361 319.90	373 341.40	386 219.20	393 374.50	382 823.90	395 920.00	383 789.30	388 227.20
Activos netos de las SB5	72 870.20	78 430.50	82 021.70	84 293.70	85 285.50	88 810.30	92 352.50	95 544.20	76 766.40	80 790.40	78 670.20	80 130.30

Fuente: Consar, *Información estadística. Activos netos*, disponible en: http://www.consar.gob.mx/SeriesTiempo/Series.aspx?cd=160&cdAlt=False, consultado el 21 de marzo de 2011.

a) Obtenga las proporciones para cada Siefore y el total anual de cada una de ellas.
b) Grafique los totales anuales para cada Siefore.

10. En el siguiente cuadro se presentan datos de la Consar con la composición de las carteras de inversiones de las Siefores al cierre de febrero de 2011.

Composición de las inversiones (Cifras porcentuales al cierre de febrero de 2011)								
Tipo de instrumento		**Siefore básica 1**	**Siefore básica 2**	**Siefore básica 3**	**Siefore básica 4**	**Siefore básica 5**	**Siefores adicionales**	**Total**
Renta variable nacional	Renta variable nacional	0.0	6.9	8.3	11.4	13.2	4.6	8.3
Renta variable internacional	América	0.0	5.7	8.1	10.9	10.8	1.2	7.6
	Asia	0.0	0.8	1.0	1.3	0.7	0.3	0.9
	Europa	0.0	0.8	1.0	1.2	0.7	0.3	0.9
	Oceanía	0.0	0.1	0.2	0.2	0.2	0.0	0.1
Deuda privada nacional	Alimentos	0.9	0.5	0.5	0.4	0.4	0.2	0.5
	Automotriz	0.4	0.3	0.3	0.2	0.3	0.0	0.3
	Banca de desarrollo	0.7	0.8	0.9	0.8	0.5	0.1	0.8
	Bancario	0.9	0.9	0.9	0.9	0.8	1.2	0.9
	Bebidas	0.7	0.5	0.4	0.4	0.4	0.0	0.5
	Cemento	0.2	0.3	0.3	0.3	0.3	0.1	0.3
	Centros comerciales	0.0	0.0	0.0	0.0	0.0	0.1	0.0
	Consumo	0.7	0.6	0.5	0.5	0.5	0.2	0.6
	Deuda CP	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Estados	0.7	0.6	0.6	0.6	0.8	1.8	0.6
	Europesos	2.6	2.3	2.1	2.1	1.8	0.1	2.2
	Grupos industriales	0.5	0.6	0.7	0.6	0.5	0.2	0.6
	Hoteles	0.0	0.0	0.0	0.1	0.0	0.0	0.0
	Infraestructura	0.7	0.6	0.7	0.6	0.4	0.4	0.6
	Otros	1.3	1.1	1.1	1.0	0.7	0.3	1.1
	Papel	0.5	0.4	0.4	0.3	0.3	0.0	0.4
	Paraestatal	4.3	3.5	3.1	2.9	2.6	1.1	3.2
	Servicios financieros	0.1	0.1	0.2	0.2	0.1	0.1	0.2
	Siderúrgica	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Telecom	2.1	1.8	1.7	1.6	1.5	0.6	1.7
	Transporte	0.0	0.1	0.1	0.1	0.1	0.0	0.1
	Vivienda	5.3	3.8	3.3	2.9	2.8	0.9	3.4
Estructurado	Estructurados	0.0	1.8	3.1	3.1	2.4	0.0	2.4
Deuda internacional	Deuda internacional	4.0	3.1	3.1	3.2	2.7	0.1	3.2
Deuda gubernamental	BOND182	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	BONDESD	1.1	0.6	0.1	0.1	0.3	15.5	0.4
	BONOS	15.5	22.9	23.0	20.0	26.6	13.6	21.5
	BPA182	14.5	1.4	0.9	0.6	1.0	1.4	2.3
	BPAS	0.0	0.0	0.0	0.0	0.1	0.0	0.0
	BPAT	0.0	0.2	0.0	0.0	0.0	2.1	0.1
	CBIC	1.2	4.6	6.0	6.4	3.2	0.0	5.1
	CETES	2.8	7.3	5.5	4.1	5.1	46.2	5.6
	DEPBMX	0.0	0.2	0.0	0.0	0.0	0.0	0.0

Composición de las inversiones								
(Cifras porcentuales al cierre de febrero de 2011)								
Tipo de instrumento		Siefore básica 1	Siefore básica 2	Siefore básica 3	Siefore básica 4	Siefore básica 5	Siefores adicionales	Total
Deuda gubernamental	UDIBONO	35.4	21.4	18.0	16.0	11.7	4.9	19.5
	UMS	1.2	1.7	2.8	3.6	4.6	0.7	2.7
	REPORTOS	1.7	1.9	1.3	1.6	1.9	1.8	1.6
	Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0

Fuente: Consar, *Información estadística. Inversiones de las Siefores*, disponible en: http://www.consar.gob.mx/SeriesTiempo/CuadroInicial.aspx?md=7, consultado el 28 de marzo de 2011.

Elabore una gráfica de barras para el total de las inversiones de las Siefores ordenándolas de menor a mayor.

capítulo 3

Medidas

Sumario

3.1 Medidas de posición
- 3.1.1 Media aritmética
- 3.1.2 Media ponderada
- 3.1.3 Media armónica
- 3.1.4 Media geométrica
- 3.1.5 Medias o promedios móviles
- 3.1.6 Mediana
- 3.1.7 Moda
- 3.1.8 Percentiles
- 3.1.9 Relación entre la media, la mediana y la moda

3.2 Medidas de dispersión
- 3.2.1 Rango
- 3.2.2 Desviación media
- 3.2.3 Desviación intercuartílica
- 3.2.4 Varianza y desviación estándar
- 3.2.5 Aplicaciones comunes de la desviación estándar

3.3 Medidas de composición: la proporción

3.4 Medidas de forma: momentos
- 3.4.1 Tercer momento respecto a la media y el coeficiente de sesgo
- 3.4.2 Cuarto momento respecto a la media y el coeficiente de curtosis o apuntamiento

3.5 Funciones estadísticas de Excel y el complemento "Análisis de datos"
- 3.5.1 Estadística descriptiva
- 3.5.2 Media móvil
- 3.5.3 Percentil y jerarquía

3.6 Resumen

3.7 Fórmulas del capítulo

3.8 Ejercicios adicionales

El tema que trata acerca de las medidas es de gran importancia en estadística y es sumamente útil en la práctica. Por lo general, cualquier adolescente o adulto sabe, cuando menos intuitivamente, qué es una media (o promedio) y una proporción; basta con preguntar: ¿cuál es el promedio de las edades de todos tus hermanos y primos?, para obtener una respuesta intuitiva, si no es que calculada, y considerablemente precisa de cuál es ese promedio; también es muy probable que sepa cómo calcular de forma exacta ese promedio. Lo mismo sucede con la proporción: si se pregunta, a la vista de un conjunto cualquiera de personas, cuál es la proporción de hombres o mujeres en él, se produce fácilmente una respuesta, de nuevo aproximada o calculada, considerablemente precisa y, asimismo, es común que la persona que responde sepa cómo se calcula esa proporción.

La media aritmética o promedio aritmético es la más importante de las medidas que en estadística se clasifican como *medidas de posición*, también conocidas como *medidas de tendencia central* y entre las que también se encuentran la mediana y la moda, además de otras que se revisan en la sección 3.1 de este capítulo.

Aunque no suele clasificarse de esta manera, en este texto se propone clasificar la proporción, dada su importancia, como una medida de composición y se revisa en la sección 3.3. Además de estas medidas de posición y de composición, en la sección 3.2 "Medidas de dispersión" se estudian las medidas de dispersión, entre las que se encuentran las que suelen ser las más conocidas: la varianza y la desviación estándar. Por otra parte, en la sección 3.4 se estudian lo que se clasifica como medidas de forma: los momentos de una distribución respecto a su media y que permiten analizar, precisamente, la forma de una distribución, a través de 2 medidas: por un lado, el coeficiente de sesgo para revisar si un conjunto de datos es simétrico o sesgado hacia uno de sus lados y, por otra parte, el coeficiente de curtosis, que mide qué tan aplanada o qué tan apuntada es una distribución.

Al igual que en otros capítulos se intercalan secciones que tratan acerca de cómo calcular muchas de esas medidas mediante funciones de Excel.

3.1 Medidas de posición

Como se mencionaba, a las medidas de posición se les conoce también como *medidas de tendencia central* aunque consideramos que, dada la agrupación que aquí se hace, es más apropiado llamarlas *medidas de po-*

sición porque, aun cuando la media y la mediana que se estudian aquí sí se refieren a posiciones centrales de una distribución, la moda no necesariamente se refiere al "centro" de una distribución y los percentiles (en particular los cuartiles 1 y 3) definitivamente no se refieren a posiciones centrales. En las subsecciones correspondientes se revisan los detalles.

3.1.1 Media aritmética

La **media aritmética**, también conocida como *promedio*, es una medida que prácticamente todas las personas conocen, al menos a nivel intuitivo, y se calcula sumando el total de los datos o valores de la variable para luego dividir esa suma entre el número de datos sumados. A esa medida común se le llama indistintamente *media, promedio* o *media aritmética*. Por ejemplo, si se tienen 5 personas cuyas edades son 15, 18, 25, 33 y 64 años, la media aritmética o promedio de edades de este grupo sería:

Media aritmética o **promedio.** Medida que se calcula sumando el total de los datos o valores de la variable para luego dividir esa suma entre el número de datos sumados.

$$\text{Media aritmética o promedio} = \frac{15 + 18 + 25 + 33 + 64}{5} = \frac{155}{5} = 31$$

El resumen de este procedimiento para calcular la media se puede plantear mediante la siguiente fórmula:

$$\bar{X} = \frac{\sum X_i}{n} \tag{3.1}$$

La forma completa de escribir esta fórmula es:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

que se lee como "la media es igual a la sumatoria desde i igual a 1 hasta n, de las observaciones X_i". Sin embargo, en lo sucesivo se simplifica la notación del operador de la sumatoria y se le quitan el subíndice y el superíndice, como en la fórmula inicial, ya que por lo general, estas expresiones se entienden adecuadamente sin esos índices.

En el siguiente ejemplo se ilustran los procedimientos para calcular esta media en los 3 tipos de tablas que se revisaron antes: tablas simples y tablas de datos agrupados (de datos y frecuencias y de clases y frecuencias).

EJEMPLO 3.1

La media en una serie simple

En la tabla 3.1 se muestran las edades de un grupo de 60 padres de familia. Calcular la media.

Tabla 3.1 Edades de un grupo de 60 padres de familia

55	58	78	52	54	63
52	52	64	56	58	78
54	65	58	56	50	52
53	58	49	53	63	52
50	54	54	56	63	48
63	68	55	57	49	53
68	54	46	75	56	65
54	50	65	64	65	65
54	50	52	65	68	63
55	55	58	65	52	70

Solución: La suma de esos 60 valores es de 3 487, por lo que la media es de:

$$\bar{X} = \frac{\sum X_i}{n} = \frac{3\,487}{60} = 58.1166 \approx 58.12$$

EXCEL Si se agrupan estos datos en las primeras columnas y renglones de una hoja de Excel, en las celdas A1:F10 y se anota en cualquier otra celda "=PROMEDIO(A1:F10)", se obtiene el mismo resultado anterior: 58.1166.

Revisando esa anotación "=PROMEDIO(A1:F10)" se puede fácilmente ver cómo se utilizan las numerosas funciones que ofrece ese paquete de Microsoft.

En primer lugar, vale la pena destacar que las comillas no se anotan; aquí se colocan para separar esa mención del resto del texto. Así, el signo de igual, que sería la primera anotación en la celda de Excel, le permite identificar al programa si lo que sigue es una operación o una función, que es el caso aquí. La parte restante es propiamente la función, en donde "Promedio" es el nombre de la función y "(A1:F10)" es el argumento de la función que, en este caso, le indican al programa el rango de celdas en donde se encuentran los datos para los que se solicita calcular el promedio.

El formato de todas las funciones de Excel es el mismo: NOMBREDELAFUNCIÓN(Argumentos) y, en caso de cualquier duda, se puede acudir a la sección de ayuda que explica detalladamente cada función, a la que se accede oprimiendo F1 del teclado de la computadora.

EJEMPLO 3.2

La media en una serie de datos y frecuencias

Agrupar los datos de edades en una tabla de datos y frecuencias y calcular la media aritmética.

Solución:

Los datos se agrupan en las 2 primeras columnas de la tabla 3.2.

Tabla 3.2 Edades de 60 padres de familia agrupados en datos y frecuencias

x (edad)	*f*	*f*(*x*)
46	1	46
48	1	48
49	2	98
50	4	200
52	7	364
53	3	159
54	7	378
55	4	220
56	4	224
57	1	57
58	5	290
63	5	315
64	2	128
65	7	455
68	3	204
70	1	70
75	1	75
78	2	156
	60	3 487

Ahora, como el cálculo de la media aritmética requiere que se sumen todos los datos, si éstos se agrupan de acuerdo con su frecuencia, sumarlos equivale a sumar los productos de cada dato (x) por la frecuencia correspondiente (f), ya que, por ejemplo, como el valor 50 aparece 4 veces, es lo mismo sumar 4 veces 50 que multiplicar 50 por 4. Se resumen estas operaciones en la columna con encabezado "$f(x)$".

Nótese que, por supuesto, la suma de las frecuencias es 60, el número de padres de familia y que la suma de los productos $f(x)$ es 3 487, el mismo valor que se encontró antes para la suma de los 60 valores en la serie simple.

En seguida resumimos el procedimiento:

- Multiplicar cada dato por la frecuencia correspondiente: $f(x)$
- Sumar estos productos: $\Sigma f(x)$
- Dividir esta suma entre el número de elementos, o sea la suma de las frecuencias: Σf. Nótese, además, que la suma de las frecuencias es igual al número de elementos o, en símbolos: $\Sigma f_i = n$.

A su vez, este procedimiento se puede resumir en la siguiente fórmula:

$$\bar{X} = \frac{\sum f_i X_i}{\sum f_i} = \frac{\sum f_i X_i}{n} \qquad \textbf{(3.2)}$$

En donde,

$\Sigma f_i X_i$ es la suma de los productos de cada valor por su frecuencia, y

Σf_i es la suma de las frecuencias que es, a la vez, el número de elementos de la serie, el número de padres de familia. Esto último en símbolos: $\Sigma f_i = n$.

Entonces, la media de esta serie de datos y frecuencias es:

$$\bar{X} = \frac{\sum f_i X_i}{\sum f_i} = \frac{3\,487}{60} = 58.12$$

la cual es el mismo resultado que se obtuvo antes. A partir de aquí se deja de utilizar el símbolo ≈, que significa "aproximadamente igual a" y que se utiliza para marcar cifras redondeadas.

EXCEL Aunque no existe una función específica para calcular medias en distribuciones de datos y frecuencias, es clara la ayuda que el paquete puede prestar en la realización de los cálculos de la tabla.

EJEMPLO 3.3

La media en una serie de clases y frecuencias

Convertir los datos de las edades de los 60 padres de familia en una serie de clases y frecuencias y calcular la media aritmética a partir de esta serie.

Solución:

En las 2 primeras columnas de la tabla 3.3 se muestran los datos de las edades, agrupadas en una serie de clases y frecuencias.

Tabla 3.3 Agrupamiento de las edades de 60 padres de familia en una serie de clases y frecuencias (columnas "*x*" y "*f*") y cálculos para determinar la media

x	*f*	Pm	*f* (Pm)
45 a menos de 50	4	47.5	190
50 a menos de 55	21	52.5	1 102.5
55 a menos de 60	14	57.5	805
60 a menos de 65	7	62.5	437.5
65 a menos de 70	10	67.5	675
70 a menos de 75	1	72.5	72.5
75 a menos de 80	3	77.5	232.5
Totales	60		3 515

En el caso de una serie de clases y frecuencias, se pierde cierta información al ubicar los valores dentro de rangos (las clases), por lo que ya no se tienen los valores exactos, como se mencionó en el capítulo anterior. Para utilizar estos datos se emplea un valor para representar a todos los que se encuentran en cada clase. Este valor que representa a todos los de su clase es el **punto medio de clase** o el **valor central de cada clase** el cual es, precisamente, el promedio entre los 2 límites de cada clase.

Por ejemplo, la primera clase va de 45 a menos de 50, por lo que su punto medio es $\frac{(45 + 50)}{2} = 47.5$. Los demás puntos medios de clase que se muestran en la tabla 3.3 se calcularon de la misma manera. El resto del procedimiento para calcular la media de una serie de clases y frecuencias es igual al que se siguió antes con los valores agrupados en serie de datos y frecuencias. En resumen, el procedimiento para el cálculo para una serie de clases y frecuencias consiste en:

- Determinar el punto medio de cada clase, *Pm*;
- Multiplicar cada punto medio de clase por la frecuencia correspondiente, $f(Pm)$;
- Sumar estos productos, $\Sigma f(Pm)$;
- Dividir esta suma entre el número de elementos, o sea la suma de las frecuencias Σf.

Para resumir este procedimiento se utiliza la siguiente fórmula:

$$\bar{X} = \frac{\sum f_i Pm_i}{\sum f_i} \tag{3.3}$$

Y, con los datos de la tabla 3.3, se tiene:

$$\bar{X} = \frac{\sum f_i Pm_i}{\sum f_i} = \frac{3\,515}{60} = 58.58$$

Un detalle que es importante observar aquí es que este resultado de la media, 58.58, no es igual al que se encontró para las tablas simple y de datos y frecuencias, 58.12, y la razón es que, como se comentó, el agrupamiento de los datos en clases hace que se pierda información. Esto último se refleja en el uso del punto medio de clase para representar a todos los elementos de cada clase.

Sin embargo, por otro lado, esta pérdida de precisión en el cálculo de medidas a partir de series de clases y frecuencias comúnmente se compensa por la sencillez en el manejo de datos agrupados de esta manera. Cuando se tienen grandes cantidades de datos, una serie de clases y frecuencias es la mejor manera de agruparlos.

Punto medio de clase o **valor central de cada clase.** Valor que representa a todos los de su clase.

EXCEL Aunque no existe una función específica para calcular medias en distribuciones de clases y frecuencias, debe resultar clara la ayuda que el paquete puede prestar en la realización de los cálculos de la tabla.

3.1.2 Media ponderada

La **media ponderada** se utiliza principalmente para darle un peso relativo diferente a cada uno de los valores de la variable. Se puede ilustrar esto con datos de precios de las acciones que cotizan en bolsa de valores.

Media ponderada. Se utiliza principalmente para darle un peso relativo diferente a cada uno de los valores de la variable.

EJEMPLO 3.4

En la tabla 3.4 se presentan datos de precios de 10 acciones que cotizan en una bolsa de valores, junto con el volumen negociado (número de acciones negociadas) a determinada fecha y se incluye una columna final con el producto del precio por el volumen.

Tabla 3.4 Precios de cierre y volumen negociado de 10 acciones bursátiles

Clave de la acción	Precio de cierre (C)	Volumen negociado (V)	C × V
CEL	28.44	1 500	42 660
CEMEXCPO	53.51	4 946 300	264 676 513
CIEB	22.99	65 500	1 505 845
COMERCIUBC	15.16	513 600	7 786 176
CONTAL	17.99	866 200	15 582 938
DESCB	2.8	116 100	325 080
ELEKTRACPO	75.95	102 000	7 746 900
FEMSAUBD	74.76	561 900	42 007 644
GCARSOA1	22.16	576 800	12 781 888
GEOB	30.2	1 064 800	32 156 960
Suma	343.96	8 814 700	384 612 604

Si se calcula la media aritmética de los precios se tiene:

$$\bar{X} = \frac{\sum X_i}{n} = \frac{343.96}{10} = 34.40$$

Sin embargo, este promedio no toma en cuenta las diferencias en el número de acciones de cada empresa que se negociaron ese día. Por ejemplo, se negociaron solamente 1 500 acciones de CEL, pero se negociaron 4 946 300 de CEMEXCPO y, por eso, es posible que se desee dar un mayor peso al precio de estas acciones al calcular el promedio. Esto es lo que hace un promedio ponderado.

Lo que se hace es multiplicar cada precio por el volumen de acciones negociadas, con lo cual este volumen es el que da el peso relativo a cada precio: a mayor volumen negociado, mayor peso (ponderación) del precio en el promedio. Después, se divide la suma de estos productos entre la suma de los volúmenes. Este procedimiento se puede resumir en la siguiente fórmula:

$$\bar{X}_p = \frac{\sum p_i X_i}{\sum p_i} \tag{3.4}$$

En donde

X_i es la variable de interés, en este caso, los precios, y

P_i son los pesos o ponderaciones, en este caso, los volúmenes negociados.

Así, en el ejemplo, la media ponderada es:

$$\bar{X}_p = \frac{\sum p_i X_i}{\sum p_i} = \frac{384\ 612\ 604}{8\ 814\ 700} = 43.63$$

Se nota que el promedio de precios subió principalmente debido al mayor peso que se le dio al precio de, precisamente, CEMEXCPO, que tiene uno de los más altos.

Incidentalmente, este proceso de ponderación, aunque con algunas consideraciones adicionales, es el usado para calcular 2 índices muy conocidos: los índices de precios al consumidor, que se utilizan para medir la inflación de los precios de una canasta de artículos básicos y los índices bursátiles como el Índice de Precios y Cozaciones de la Bolsa Mexicana de Valores, el DJIA (*Down Jones Industrial Average* o Índice Industrial Dow Jones) de la Bolsa de Valores de Nueva York, o el BSVP, el índice de precios de acciones de la Bolsa de Valores de São Paulo. Se revisa este tema con mayor detalle en el capítulo 15.

EXCEL No existe función específica para calcular medias ponderadas.

3.1.3 Media armónica

Media armónica. Es el recíproco de la media aritmética de los recíprocos de los valores individuales.

La **media armónica** es el recíproco de la media aritmética de los recíprocos de los valores individuales, o en símbolos:

$$\bar{X}_a = \frac{1}{\dfrac{\sum_{i=1}^{n} \frac{1}{X_i}}{n}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{X_i}} \tag{3.5}$$

Se le suele utilizar para promediar distintas velocidades desarrolladas en distancias iguales.

Ejemplo 3.5

Supóngase que en una carrera de automóviles de relevos, 3 pilotos condujeron 3 000 kilómetros, es decir 1 000 kilómetros cada uno, con las siguientes velocidades promedio:

Conductor	Velocidad promedio (km/h)
A	100
B	120
C	140

La media aritmética de estas velocidades es

$$\frac{100 + 120 + 140}{3} = 120$$

La media armónica es

$$MA = \frac{3}{\frac{1}{100} + \frac{1}{120} + \frac{1}{140}} = \frac{3}{0.02547619} = 117.757$$

En este ejemplo no es correcto usar la media aritmética porque eso supondría que los conductores manejaron el mismo número de horas, lo cual no es el caso, evidentemente, porque el conductor más veloz manejó durante menos tiempo y viceversa. El promedio correcto es la media armónica, 117.757 kilómetros por hora.

En otras palabras, si el promedio de velocidad fuera la media aritmética de 120 km/h entonces la distancia recorrida en total por los 3 pilotos sería:

Conductor A: 1 000/100 = 10 horas de manejo
Conductor B: 1 000/120 = 8.3333 horas de manejo, y
Conductor C: 1 000/140 = 7.1428 horas de manejo, para un total de 25.4761 horas.

En este tiempo y a una velocidad promedio de 120 km/h se recorrería un total de 120 × 25.4761 = 3 057 kilómetros, lo cual no fue lo supuesto.

Por otro lado, a una velocidad de 117.757, la distancia recorrida en 25.4761 horas serían los 3 000 kilómetros planteados: 25.4761 × 117.757 = 2 999.989 y la pequeña diferencia se debe a redondeo.

Excel La función MEDIA.ARMO(100,120,140) da como resultado 117.757.

Ejemplo 3.6

Otro ejemplo de media armónica

Una fábrica de juguetes se asignó a un grupo de 4 trabajadores para completar una orden de 700 juguetes y la productividad de cada trabajador es diferente según se muestra en la siguiente tabla:

Trabajador	A	B	C	D
Productividad	10 min × juguete	6 min × juguete	15 min × juguete	4 min × juguete

a) Determine el número total de minutos que se requiere para completar la orden.

Solución:

Si a cada trabajador se le asigna el mismo número de juguetes, es decir 175 juguetes (700 ÷ 4 = 175) se tendría que:

Trabajador	Tiempo requerido para producir 175 juguetes
A	10 × 175 = 1 750 min
B	6 × 175 = 1 050 min
C	15 × 175 = 2 625 min
D	4 × 175 = 700 min
Total	6 125 min

Además se puede calcular una media del número de juguetes que cada trabajador produce, ponderada por el tiempo que cada uno de ellos se tarda en fabricarlos:

$$\bar{X}_p = \frac{6\ 125}{700} = 8.75\,\text{min} \cdot \text{juguete}$$

Sin embargo, como lo que interesa es saber cuánto tiempo se requiere para cubrir la orden aprovechando al máximo la productividad de los trabajadores, se utiliza la media armónica con los siguientes resultados:

$$\bar{X}_a = \frac{4}{\frac{1}{10} + \frac{1}{6} + \frac{1}{15} + \frac{1}{4}} = \frac{4}{0.58333333} = 6.857$$

O sea que se pueden fabricar, en promedio, 6.857 por minuto y, para completar la orden de 700 juguetes se requieren 6.85 × 700 = 4 800 min.

Excel La función MEDIA.ARMO(10,6,15,4) da como resultado 6.857.

3.1.4 Media geométrica

Media geométrica. Raíz *n*-ésima del producto de los *n* datos o valores de la variable.

La **media geométrica** es particularmente apropiada para promediar porcentajes o medidas que aumentan a una razón constante y se define como la raíz *n*-ésima del producto de los *n* datos o valores de la variable o, en símbolos:

$$\bar{X}_g = \sqrt[n]{X_1 \cdot X_2 \cdot \ldots \cdot X_n} = (X_1 \cdot X_2 \cdot \ldots X_n)^{\frac{1}{n}} \qquad (3.6)$$

EJEMPLO 3.7

Una inversión de $100 crece a razón del 5% anual durante 7 años, de manera que su valor al final de cada uno de estos 7 años sería:

Fin del año	Valor	
1	100(1.05) =	105.00
2	105(1.05) =	110.25
3	110.25(1.05) =	115.76
4	115.76(1.05) =	121.55
5	121.55(1.05) =	127.63
6	127.63(1.05) =	134.01
7	134.01(1.05) =	140.71
Suma		854.91

El promedio aritmético de estos 7 valores es:

$$\bar{X} = \frac{854.91}{7} = 122.13$$

Pero éste no es el valor del pago central, el del cuarto año. Pero, si se calcula la media geométrica:

$$\bar{X}_g = (105 \cdot 110.25 \cdot 115.76 \cdot 121.55 \cdot 127.63 \cdot 134.01 \cdot 140.71)^{\frac{1}{7}}$$

$$=121.55$$

Que es el valor de los $100 a la mitad del periodo de 7 años.

EXCEL La función =MEDIA.GEOM(105,110.25,115.76,121.55,127.63,134.01,140.71) da como resultado 121.55.

EJEMPLO 3.8

Otro ejemplo de media geométrica:

En la tabla 3.5 se muestran los datos de las ventas anuales de la Compañía Papelera La Concordia, S. A., de 2004 a 2008 y aparecen también las razones de las ventas de cada año respecto a las del año anterior.

Tabla 3.5 Datos del ejemplo 3.8

Año	Ventas anuales (millones de pesos)	Razón de las ventas respecto al año anterior
2004	$50	
2005	$36	0.72
2006	$57.6	1.6
2007	$51.84	0.9
2008	$103.68	2

La media geométrica es:

$$\bar{X}_g = (0.72 * 1.6 * 0.9 * 2)^{\frac{1}{4}} = 1.2$$

Por otra parte, la media aritmética de las razones es:

$$\bar{X} = \frac{0.72 + 1.6 + 0.9 + 2}{4} = \frac{5.22}{4} = 1.305$$

Ahora se comparan en la tabla 3.6 las ventas basadas en los 2 promedios recién calculados.

Tabla 3.6 Comparación de las ventas calculadas mediante las medias aritmética y geométrica

Año	Ventas anuales (millones de pesos)	Ventas basadas en la media geométrica (1.2)	Ventas basadas en la media aritmética (1.305)
2004	$50		
2005	$36	$60 (50 * 1.2)	$65.25 (60 * 1.305)
2006	$57.6	$72 (60 * 1.2)	$85.15125 (65.25 * 1.305)
2007	$51.84	$86.4 (72 * 1.2)	$111.122 (85.151 * 1.305)
2008	$103.68	$103.68 (86.4 * 1.2)	$145.01 (111.122 * 1.305)

Como puede verse, la media geométrica arroja el valor correcto de las ventas del último año, no la media aritmética.

EXCEL La función =MEDIA.GEOM(0.72,1.6,0.9,2) da como resultado 1.2.

3.1.5 Medias o promedios móviles

Los promedios móviles se suelen utilizar para suavizar los movimientos de series de tiempo. Para ilustrar esto, en la figura 3.1 se muestran los precios de una acción que cotiza en bolsa de valores y su promedio móvil de 10 días. La línea con muchos quiebres muestra los precios de cierre de cada día hábil y la línea continua es el promedio móvil de esos precios. Como puede verse, la línea de los promedios móviles sigue de cerca los movimientos de los precios pero en forma suavizada.

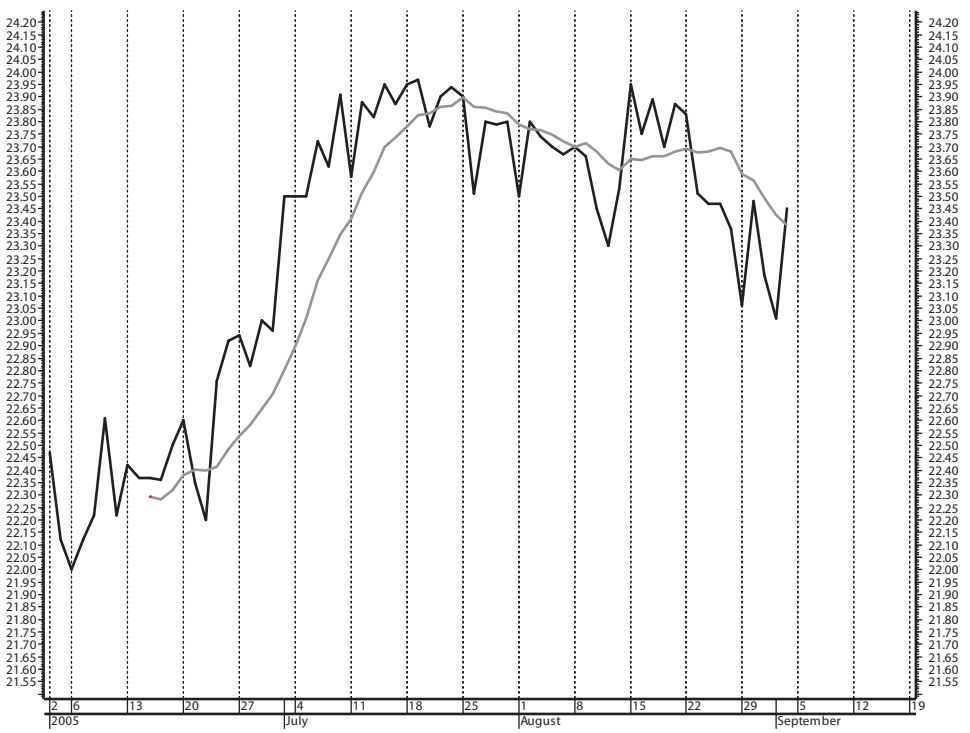


Figura 3.1 Precios de una acción bursátil y su promedio móvil de 10 días.

Los **promedios móviles** son promedios que se determinan para cada punto de la serie, calculando el promedio para el valor del punto en cuestión, más determinado número de puntos anteriores; en otras palabras son promedios que se calculan sucesivamente (es decir, en este caso, día tras día) añadiendo cada vez el nuevo día y eliminando el día más antiguo. Por ejemplo, en los promedios móviles que se muestran en la gráfica, el valor del promedio móvil del primer día fue el 10 de agosto, se calculó con el precio de la acción de ese día y los 9 precios anteriores. El promedio móvil del 11 agosto se calculó con el precio de dicho día y los 9 días anteriores, y así sucesivamente.

Promedios móviles. Promedios que se calculan sucesivamente añadiendo cada vez el nuevo día y eliminando el más antiguo.

EXCEL Se ilustra en el ejemplo siguiente la forma en la que se pueden calcular promedios móviles con este programa. El procedimiento básico consiste en plantear el PROMEDIO con el número de días que se requieren y, después, copiar la fórmula para los demás renglones de la tabla.

Nota: Para hacer esto rápidamente basta con seleccionar la celda C5, con lo que aparece un pequeño cuadro en la esquina inferior derecha de la celda. Ya que aparece este pequeño cuadro, basta con darle doble clic para que el paquete llene los demás promedios móviles.

EJEMPLO 3.9

En las 2 primeras columnas de la tabla 3.7 se presentan los flujos netos de Pemex, de agosto de 2008 a febrero de 2011, según el Banco de México y, en la tercera columna, están los promedios móviles de 5 meses.

Tabla 3.7 Flujos netos de Pemex

Ago. 2008	2 659.5	
Sep. 2008	3 041.3	
Oct. 2008	2 541.5	
Nov. 2008	4.6	
Dic. 2008	817.7	1 812.9
Ene. 2009	–703.3	1 140.4
Feb. 2009	3 560.2	1 244.1
Mar. 2009	–141.2	707.6
Abr. 2009	690.4	844.8
Mayo 2009	623.2	805.9
Jun. 2009	778.4	1 102.2
Jul. 2009	1 388.2	667.8
Ago. 2009	912.5	878.5
Sep. 2009	2 438.3	1 228.1
Oct. 2009	1 406.7	1 384.8
Nov. 2009	870.6	1 403.3
Dic. 2009	–295.1	1 066.6
Ene. 2010	1 014.7	1 087.0
Feb. 2010	1 870.8	973.5
Mar. 2010	1 852.4	1 062.7
Abr. 2010	475.4	983.6
Mayo 2010	659.4	1 174.5
Jun. 2010	338.5	1 039.3
Jul. 2010	1 900.4	1 045.2
Ago. 2010	1 682.5	1 011.2
Sep. 2010	1 888.7	1 293.9
Oct. 2010	–973.2	967.4
Nov. 2010	2 202.4	1 340.2
Dic. 2010	3 124.7	1 585.0
Ene. 2011	951.4	1 438.8
Feb. 2011	1 473.0	1 355.7

Fuente: Banco de México, disponible en: http://www.banxico.gob.mx/PortalesEspecializados/tiposCambio/estadisticas/saldosyflujosact.html, consultado el 25 de marzo de 2011.

El procedimiento para calcular los promedios móviles consiste en calcular un primer promedio de los 5 primeros valores. Si se considera que los datos están en las columnas A y B y desde el renglón 1, se puede anotar "PROMEDIO(B1:B5)" en la celda C5, con lo que se obtiene el primer valor de 1 812.9 y, después, simplemente se copia esta fórmula al resto de los renglones.

En la figura 3.2 se muestra la gráfica de ambas series y se vuelve a observar la suavización que se obtiene con los promedios móviles.

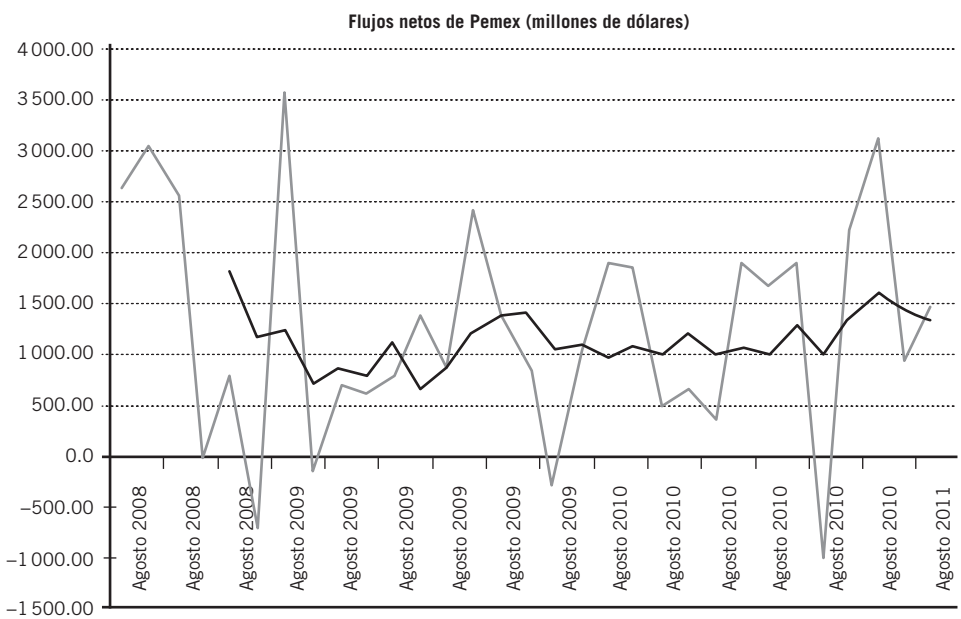


Figura 3.2 Flujos netos de Pemex con su promedio móvil de 5 meses.

3.1.6 Mediana

Mediana. Valor que ocupa el lugar central en una serie ordenada.

La **mediana** es el valor que ocupa el lugar central en una serie ordenada. El procedimiento para encontrar la mediana es el siguiente:

- Ordenar la serie, normal, aunque no necesariamente, de menor a mayor.
- Determinar la posición central, ésta se encuentra dividiendo entre 2 el número de elementos más 1 o, en símbolos, $\frac{(n+1)}{2}$.
- Identificar el valor que ocupa esta posición central, que es la mediana.

Ejemplo 3.10

Encontrar la mediana de las tasas mensuales de interés de Cetes, que se reproducen en las 2 primeras columnas de la tabla 3.8.

Tabla 3.8 Tasas de interés promedio mensual para los Certificados de la Tesorería de la Federación (Cetes) a 28 días

Datos en orden cronológico		Datos ordenados por tasa	
Mes	Tasa	Mes	Tasa
Ago. 2009	4.49	Nov. 2010	3.97
Sep. 2009	4.48	Oct. 2010	4.03
Oct. 2009	4.51	Feb. 2011	4.04
Nov. 2009	4.51	Ene. 2011	4.14
Dic. 2009	4.50	Dic. 2010	4.30
Ene. 2010	4.49	Sep. 2010	4.43
Feb 2010	4.49	Abr. 2010	4.44
Mar. 2010	4.45	Mar. 2010	4.45
Abr. 2010	4.44	Sep. 2009	4.48
Mayo 2010	4.52	Ago. 2009	4.49
Jun. 2010	4.59	Ene. 2010	4.49
Jul. 2010	4.60	Feb. 2010	4.49
Ago. 2010	4.52	Dic. 2009	4.50
Sep. 2010	4.43	Oct. 2009	4.51
Oct. 2010	4.03	Nov. 2009	4.51
Nov. 2010	3.97	Mayo 2010	4.52
Dic. 2010	4.30	Ago. 2010	4.52
Ene. 2011	4.14	Jun. 2010	4.59
Feb. 2011	4.04	Jul. 2010	4.60

Fuente: Banco de México, *Tasas y precios de referencia*, disponible en: http://www.banxico.org.mx/politica-monetaria-e-inflacion/estadisticas/otros-indicadores/tasas-interes-precios.html, consultado el 25 de marzo de 2010.

Solución:

En primer lugar, en las columnas tercera y cuarta de la tabla 3.8 se reprodujeron los datos de las tasas pero ahora en orden ascendente de su valor y ya no por meses. Como $n = 19$, $\frac{(n+1)}{2} = 10$, de manera que la tasa en la décima posición es la mediana:

$$\text{Med} = 4.49$$

Excel La función que calcula la mediana es precisamente la MEDIANA. Por ello, si se copia la tabla 3.8 a una hoja de Excel, los datos de las tasas quedarían en las columnas B (los datos originales) y D (los datos ordenados), y en los renglones del 3 al 21. Las funciones "MEDIANA(B3:B21)" y "MEDIANA(D3:D21)", producen el mismo resultado de 4.49, lo cual significa que, en Excel, no es necesario ordenar los datos antes de aplicar la función, el paquete mismo los ordena.

Se ve en el ejemplo siguiente la forma en la que se determina la mediana cuando el número de elementos es par.

Ejemplo 3.11

En la tabla 3.9 se reproducen los datos de la tabla 3.1, que son las edades de 60 padres de familia. Encontrar la mediana

Tabla 3.9 Edades de un grupo de 60 padres de familia

55	58	78	52	54	63
52	52	64	56	58	78
54	65	58	56	50	52
53	58	49	53	63	52
50	54	54	56	63	48
63	68	55	57	49	53
68	54	46	75	56	65
54	50	65	64	65	65
54	50	52	65	68	63
55	55	58	65	52	70

Solución:

En la tabla 3.10 se muestran estos mismos datos de edades, pero ahora ordenados de menor a mayor. Se muestran, además, en esta tabla números consecutivos que muestran la posición que cada edad ocupa en la serie ordenada.

Tabla 3.10 Edades de los 60 padres de familia, ordenadas y con identificación del lugar que cada una de ellas ocupa en la serie

Núm.	Edad	Núm.	Edad	Núm.	Edad	Núm.	Edad	Núm.	Edad	Núm.	Edad
1	46	11	52	21	54	31	56	41	63	51	65
2	48	12	52	22	54	32	56	42	63	52	65
3	49	13	52	23	54	33	56	43	63	53	65
4	49	14	52	24	54	34	57	44	63	54	68
5	50	15	52	25	54	35	58	45	64	55	68
6	50	16	53	26	55	36	58	46	64	56	68
7	50	17	53	27	55	37	58	47	65	57	70
8	50	18	53	28	55	38	58	48	65	58	75
9	52	19	54	29	55	39	58	49	65	59	78
10	52	20	54	30	56	40	63	50	65	60	78

La posición central de la serie es $\frac{(n+1)}{2} = \frac{(60+1)}{2} = 30.5$, por lo que la posición central la ocupan los valores en los lugares 30 y 31 y, entonces, la mediana es el promedio de los valores que ocupan esas 2 posiciones. Pero, además, como ambos valores son iguales a 56, entonces su promedio es también 56. La mediana es, entonces:

$$\text{Med} = 56$$

Excel Si se colocan los datos (no importa si están ordenados o no) en una hoja de Excel, ocuparían las celdas de la A1 a la F10, por lo que la función adecuada sería:
=MEDIANA(A1:F10), la cual produce el resultado de 56, que es el correcto.

Ejemplo 3.12

La mediana en una serie de datos y frecuencias

Se muestran en la tabla 3.11 los mismos datos de las edades, pero ahora agrupados en una serie de datos y frecuencias y es una reproducción de la tabla 3.2 que se utilizó antes para calcular la media. Encontrar la mediana.

Tabla 3.11 Edades de 60 padres de familia agrupados en datos y frecuencias

X (edades)	*f*	*f* acumulada
46	1	1
48	1	2
49	2	4
50	4	8
52	7	15
53	3	18
54	7	25
55	4	29
56	4	33
57	1	34
58	5	39
63	5	44
64	2	46
65	7	53
68	3	56
70	1	57
75	1	58
78	2	60
	60	

Aquí, al igual que antes, la mediana es el valor que ocupa la posición intermedia en la serie ordenada y esta posición se determina de la misma manera que antes

$$\text{Posición de la mediana} = \frac{n+1}{2} = \frac{60+1}{2} = 30.5$$

En esta tabla 3.11 se puede observar en la columna de la frecuencia acumulada que la posición 30.5 está en el valor X = 56 = Med, ya que el dato se repite en las posiciones 30, 31, 32 y 33.

Excel El paquete no tiene una función específica para calcular la mediana en una serie de datos y frecuencias.

Ejemplo 3.13

La mediana en una serie de clases y frecuencias

En la tabla 3.12 se reproduce la tabla 3.3 que contiene los datos de las edades agrupados en una serie de clases y frecuencias. Encontrar la mediana.

Tabla 3.12 Las edades de 60 padres de familia agrupadas en una serie de clases y frecuencias

X	*f*	*f* acumulada
45 a menos de 50	4	4
50 a menos de 55	21	25
55 a menos de 60	14	39
60 a menos de 65	7	46
65 a menos de 70	10	56
70 a menos de 75	1	57
75 a menos de 80	3	60
Totales	60	

Solución:

De nueva cuenta, la mediana es el valor que ocupa la posición intermedia en la serie ordenada y se sabe que esta posición es la 30.5. Una manera sencilla de determinar la mediana en esta serie es simplemente tomar el punto medio de la clase que contiene la mediana. Como la posición 30.5 está en la clase de 55 a menos de 60, la mediana sería el punto medio de esta clase, o sea: $\frac{(55+60)}{2} = 57.5$.

Como se ha mencionado, al agrupar los datos en clases se pierde un poco de información porque ya no se sabe el valor exacto de cada dato y por eso se utiliza el punto medio de clase para representar a cada dato en cada clase. Sin embargo, se puede hacer una mejor aproximación de la mediana a través de interpolación, representando de la siguiente manera la posición 30.5 en la clase de 55 a menos de 60:

En esta figura, el 25 que está arriba del 55 es la frecuencia acumulada hasta antes de la clase que contiene la mediana y que empie-

za en 55 y el 39 es la frecuencia acumulada hasta (incluyendo) la clase mediana.

Entonces la mediana, en la posición 30.5, está a una distancia de 30.5 – 25 = 5.5 datos (frecuencias) a partir del límite inferior de la clase. Ahora, 39 – 25 = 14 es el recorrido total de las frecuencias y $\frac{5.5}{14}$ = 0.3929 representa la proporción de la distancia que se recorre, en términos de frecuencias, desde el límite inferior. Ahora se calcula esa misma proporción pero en términos de los datos, es decir, de las edades. El recorrido de las edades es de 5 (55 – 60) y la porción 0.3929 de este recorrido de 5 es: 0.3929(5) = 1.9645. Ésta sería la distancia, en unidades de edad, recorrida desde el inicio de la clase 55 a menos de 60, por lo que la mediana así determinada es Med = 55 + 1.9645 = 56.9645, que como puede verse, es una mejor aproximación que el 57.5 que se encontró mediante el simple punto medio de clase.

Se puede resumir el procedimiento anterior de interpolación mediante la siguiente fórmula:

$$Med = L_{icmed} + i\left(\frac{\frac{n+1}{2} - f_{aacmed}}{f_{cmed}}\right) \tag{3.7}$$

En donde,

n = número de datos, que en el caso de una serie agrupada = $\sum f$.
L_{icmed} = límite inferior de la clase mediana.
i = el intervalo de clase: el límite superior de la clase menos su límite inferior.
f_{aacmed} = frecuencia acumulada hasta antes de la clase mediana.
f_{cmed} = frecuencia de la clase mediana.

Sustituyendo ahora en esta fórmula se tiene:

$$Med = L_{icmed} + i\left(\frac{\frac{n+1}{2} - f_{aacmed}}{f_{cmed}}\right) = 55 + 5\left(\frac{30.5 - 25}{14}\right) = 55 + 5(0.392857142) = 55 + 1.9645 = 56.9645$$

Que es el mismo valor que se encontró antes.

Excel El paquete no tiene una función específica para calcular la mediana en una serie de clases y frecuencias.

3.1.7 Moda

La **moda** es el valor que más se repite, es decir, el que tiene mayor frecuencia.

Moda. Valor que más se repite, es decir, el que tiene mayor frecuencia.

Ejemplo 3.14

Determinar la moda para los datos de edades de 60 padres de familia.

Solución:

A partir de la definición de moda es claro que la manera más conveniente de determinarla es a partir de una serie de datos agrupados en frecuencias. En la tabla 3.13 se reproduce la tabla 3.2, en la que se habían agrupado los valores en una tabla de datos y frecuencias.

Tabla 3.13 Edades de 60 padres de familia agrupadas en datos y frecuencias

X (edades)	*f*
46	1
48	1
49	2
50	4
52	7
53	3
54	7
55	4
56	4
57	1
58	5
63	5
64	2
65	7
68	3
70	1
75	1
78	2
	60

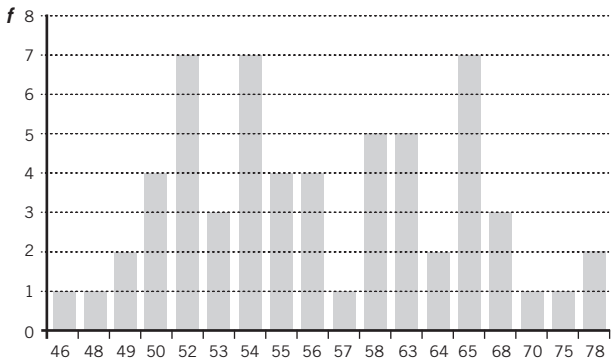


Figura 3.3 Gráfica de las edades agrupadas en una serie de datos y frecuencias.

En la tabla 3.13 se puede apreciar que la serie tiene tres modas, 52, 54 y 65, ya que los 3 valores se repiten 7 veces. Se tiene una serie con 3 modas, lo cual destaca si se grafica esta serie como en la figura 3.3.

Excel No existe una función específica en Excel para calcular la moda a partir de una serie de datos y frecuencias como la de este ejemplo.

Ejemplo 3.15

Determinar la moda para los datos de las tasas de Cetes de la tabla 3.8 que se reproduce ahora como tabla 3.14.

Tabla 3.14 Tasas de interés promedio mensual para los Certificados de la Tesorería de la Federación (Cetes) a 28 días

Mes	Tasa	Mes	Tasa	Frecuencia
Ago. 2009	4.49	Nov. 2010	3.97	1
Sep. 2009	4.48	Oct. 2010	4.03	1
Oct. 2009	4.51	Feb. 2011	4.04	1
Nov. 2009	4.51	Ene. 2011	4.14	1
Dic. 2009	4.50	Dic. 2010	4.30	1
Ene. 2010	4.49	Sep. 2010	4.43	1
Feb. 2010	4.49	Abr. 2010	4.44	1
Mar. 2010	4.45	Mar. 2010	4.45	1
Abr. 2010	4.44	Sep. 2009	4.48	1
May. 2010	4.52	Ago. 2009	4.49	3
Jun. 2010	4.59	Ene. 2010	4.49	
Jul. 2010	4.60	Feb. 2010	4.49	
Ago. 2010	4.52	Dic. 2009	4.50	1
Sep. 2010	4.43	Oct. 2009	4.51	2
Oct. 2010	4.03	Nov. 2009	4.51	
Nov. 2010	3.97	Mayo 2010	4.52	2
Dic. 2010	4.30	Ago. 2010	4.52	
Ene. 2011	4.14	Jun. 2010	4.59	1
Feb. 2011	4.04	Jul. 2010	4.60	1

Fuente: Banco de México, *Tasas y precios de referencia*, disponible en: http://www.banxico.org.mx/politica-monetaria-e-inflacion/estadisticas/otros-indicadores/tasas-interes-precios.html, consultado el 25 de marzo de 2010.

En esta tabla se nota claramente que la moda es 4.49, que es la tasa que más se repite.

Excel La función "=MODA(B2:B20)", para los datos en orden de fechas, o "=MODA(D2:D20)", en orden de magnitud, producen el mismo resultado de moda = 7.04. De lo anterior, se puede observar que, para la moda, no es necesario agrupar los datos para determinarla; la función de Excel hace todo el trabajo. Como se mencionó en el ejemplo anterior, el paquete no tiene funciones específicas para determinar la moda en datos agrupados y, como se vio antes, tampoco las tiene para la media ni la mediana.

Ejemplo 3.16

La moda en una serie de clases y frecuencias

Determinar la moda para los datos de las edades de padres de familia que hemos utilizado, los cuales se reproducen ahora como tabla 3.15 (vea la página siguiente).

Solución:
En estos datos agrupados ya sólo se tiene una clase con mayor frecuencia, la que va de 50 a menos de 55 y, por lo tanto, ya sólo se tiene una moda, a diferencia de la serie original, en la que se tenían 3 modas. Esto se debe, por supuesto, al agrupamiento de los datos y a la pérdida de información consiguiente, como ya se ha mencionado antes.

Tabla 3.15 Las edades de 60 padres de familia agrupadas en una serie de clases y frecuencias

X	f
45 a menos de 50	4
50 a menos de 55	21
55 a menos de 60	14
60 a menos de 65	7
65 a menos de 70	10
70 a menos de 75	1
75 a menos de 80	3
Totales	60

El procedimiento para determinar la moda se resume en la siguiente fórmula:

$$Mod = L_{\text{icmod}} + i\left(\frac{d_1}{d_1 + d_2}\right) \tag{3.8}$$

En donde,
L_{icmod} = límite inferior de la clase modal (la clase que contiene la moda)
i = el intervalo de clase: el límite superior de la clase menos su límite inferior
d_1 = la diferencia entre la frecuencia de la clase modal y la frecuencia de la clase anterior
d_2 = la diferencia entre la frecuencia de la clase modal y la frecuencia de la clase siguiente

Sustituyendo:

$$Mod = L_{\text{icmod}} + i\left(\frac{d_1}{d_1 + d_2}\right) = 50 + 5\left(\frac{17}{17 + 7}\right) = 50 + 5\left(\frac{17}{24}\right) = 50 + 5(0.7083) = 50 + 3.542 = 53.542$$

EXCEL El paquete no tiene función específica para determinar la moda en datos agrupados.

3.1.8 Percentiles

Ya se vio antes que la mediana es el valor que ocupa el lugar central en una serie ordenada; en otras palabras, la mediana divide a la serie en 2 mitades. Siguiendo la misma idea se puede dividir la serie en 100 partes, con lo que se tendrían percentiles. Si se divide la serie en 4 partes, se dice que se tienen cuartiles y, de la misma manera, si se divide la serie en 10 partes, entonces se tienen deciles.

Como se requiere determinar la posición que estas medidas tienen en la serie ordenada, se puede utilizar la siguiente fórmula:

$$P = (n + 1)\frac{p}{100} \tag{3.9}$$

Por ejemplo, generalizando la fórmula $P = \frac{n+1}{2}$ que se utilizó para determinar la posición de la mediana, se tiene que, $P_{Med} = (n + 1)\frac{50}{100}$ que es la misma fórmula de antes y que, por otro lado, es la misma posición que la del segundo cuartil, o el quinto decil o el quincuagésimo percentil: es la posición que divide la serie ordenada en 2.

Si se quiere encontrar la posición que divide la serie en una primera cuarta parte y la parte superior con 75% de los datos que es, precisamente, el primer cuartil, entonces:

$$P_{Q_1} = (n + 1)\frac{25}{100}$$

Y, para el tercer cuartil,

$$P_{Q_3} = (n + 1)\frac{75}{100}$$

En las secciones siguientes se presentan algunos ejemplos.

3.1.8.1 Cuartiles

EJEMPLO 3.17

En la tabla siguiente se muestran los números de acciones negociadas de cierta empresa (en miles), en 50 días de operaciones. Determinar los valores de los cuartiles.

3	10	19	27	34	38	48	56	67	74
4	12	20	29	34	39	48	59	67	74
7	14	21	31	36	43	52	62	69	76
9	15	25	31	37	45	53	63	71	79
10	17	27	34	38	47	56	64	73	80

Solución:

Como la serie tiene 50 elementos, la posición del primer cuartil es

$$P_{Q_1} = (n + 1)\frac{25}{100} = (50 + 1)(0.25) = 12.75$$

La posición del segundo cuartil (la mediana) es:

$$P_{Med} = P_{Q_2} = (n + 1)\frac{50}{100} = 51(0.5) = 25.5$$

Y, para el tercer cuartil,

$$P_{Q_S} = (n + 1)\frac{75}{100} = (51)(0.75) = 38.25$$

Además, como la serie ya se ordenó para determinar el valor de los cuartiles, se cuenta la posición que cada valor ocupa en la serie. Así, el 20 ocupa la posición 12 y el 21 ocupa la 13, por lo que el valor de la posición 12.75 es 75% del camino entre 20 y 21 que es, precisamente, 20.75. En símbolos: $Q_1 = 20.75$.

La mediana está en la posición 25.5, es decir entre la posición 25, que corresponde al valor 38, que es el mismo valor que aparece en la posición 26, por lo que $Q_2 = \text{Med} = 38$. Finalmente, para el cuartil 3, la posición 38 tiene el valor 62 y la posición 39 tiene el valor 63 por lo que la posición 38.25 vale 62.25, o $Q_3 = 62.25$.

Es muy importante no olvidar que la posición de estas medidas es distinta a su valor: 12.75 es la posición del primer cuartil, no su valor. El valor de este primer cuartil es 20.75. Resumiendo:

Medida	Posición en la serie	Valor
Primer cuartil	12.75	20.75
Segundo cuartil = mediana	25.5	38
Tercer cuartil	38.25	62.25

EXCEL Si se colocan los datos en una hoja, en las celdas A1 a J5, las siguientes funciones calculan los cuartiles 1, 2 (la mediana) y 3:

=CUARTIL(A1:J5,1) = 22
=CUARTIL(A1:J5,2) = 38
=CUARTIL(A1:J5,3) = 62.25

Como puede verse, el valor encontrado manualmente para el cuartil 1 no coincide con el que Excel determina; este autor no encontró la explicación y está en contacto con Microsoft para tratar de determinar la causa.

Los cuartiles en una serie de clases y frecuencias

El procedimiento que se sigue para determinar el valor de los cuartiles en una serie de clases y frecuencias es el mismo que se vio antes para la mediana, salvo que se sustituye en el proceso de interpolación (en la fórmula) la posición correspondiente. Así, las fórmulas son las siguientes:

$$Q_1 = L_{icQ_1} + i\left(\frac{\frac{(n+1)}{4} - f_{aacQ_1}}{f_{cQ_1}}\right) \tag{3.10}$$

$$Q_2 = Med = L_{icQ_{12}} + i\left(\frac{\frac{n+1}{2} - f_{aacQ_2}}{f_{cQ_2}}\right) \tag{3.11}$$

$$Q_3 = L_{icQ_3} + i\left(\frac{\frac{3(n+1)}{4} - f_{aacQ_3}}{f_{cQ_3}}\right) \tag{3.12}$$

EJEMPLO 3.18

En la tabla siguiente se muestra el número de pasajeros que viajaron en 100 vuelos de Aerolíneas Mayas, incluyendo la frecuencia acumulada. Determinar el valor de los cuartiles.

X Núm. de pasajeros	*f*	*f* acumulada
40 a 49	3	3
50 a 59	7	10
60 a 69	12	22
70 a 79	28	50
80 a 89	22	72
90 a 99	18	90
100 a 109	8	98
109 a 119	2	100

Solución:

La posición de los cuartiles es: 101(0.25) = 25.25 para el primero, 101(0.5) = 50.5 para el segundo, y 101(0.75) = 75.75, para el tercero, por lo que la clase del primer cuartil es la de 70 a 79, la del segundo cuartil o mediana no es una sola sino que el valor de este cuartil se encuentra en el límite entre la clase de 70 a 79 (con 50 de frecuencia acumulada) y la de 80 a 89 (que es donde empieza la frecuencia 51). La clase del tercer cuartil es la de 90 a 99. Los valores de los cuartiles son:

$$Q_1 = L_{icQ_1} + i\left(\frac{\frac{n+1}{4} - f_{aacQ_1}}{f_{cQ_1}}\right) = 70 + 10\left(\frac{25.25 - 22}{28}\right)$$

$$= 70 + 10(0.115) = 70 + 1.15 = 71.15$$

El cuartil 2 o mediana, como se encuentra entre los límites de 2 clases es el promedio de 79 y 80, o sea 79.5:

$$Q_2 = 79.5$$

$$Q_3 = L_{icQ_3} + i\left(\frac{\frac{n+1}{4} - f_{aacQ_3}}{f_{cQ_3}}\right) = 90 + 10\left(\frac{75.75 - 72}{18}\right)$$

$$= 90 + 10(0.2083) = 92.083$$

EXCEL El paquete tiene la función PERCENTIL, la cual permite calcular cualquier posición, incluyendo los cuartiles. Para la tabla anterior puede utilizarse el conjunto de datos original y, por ejemplo, la función

=PERCENTIL(RangoDeDatos,0.10)

En este caso, la opción "RangoDeDatos" especificaría las celdas de Excel en donde se encuentran los datos y el 0.10 señala que se pide calcular el primer decil, es decir, el valor que divide a los datos en 10% más bajo y 90% más alto. Si se usa 0.25, la función da el valor del primer cuartil y, si se usa 0.50, arroja el valor del cuartil 2, que es lo mismo que la mediana.

3.1.8.2 Deciles

El procedimiento para encontrar valores de deciles es similar a los explicados antes. Aquí, simplemente para ilustrar su utilidad, se reproduce en seguida una tabla sobre ingresos de familias, divididos en deciles.

Tabla 3.16 Ingreso corriente trimestral por hogar, en deciles de hogares, según el tamaño de su localidad y su coeficiente de Gini, 2008 (en miles de pesos)

			Tamaño de la localidad			
	Total		De 2 500 y más habitantes		Menos de 2 500 habitantes	
Deciles de hogares[1]	Hogares	Ingreso	Hogares	Ingreso	Hogares	Ingreso
Ingreso corriente	26 732 594	987 179 918	21 210 281	887 703 337	5 522 313	99 476 581
I	2 673 259	15 001 200	1 115 488	6 574 672	1 557 771	8 426 528
II	2 673 259	27 302 098	1 566 851	16 106 987	1 106 408	11 195 111
III	2 673 259	37 350 360	1 918 364	26 921 863	754 895	10 428 497
IV	2 673 259	47 051 141	2 101 733	37 028 369	571 526	10 022 772
V	2 673 259	58 102 095	2 255 290	49 080 578	417 969	9 021 517
VI	2 673 259	72 150 866	2 324 410	62 784 910	348 849	9 365 956
VII	2 673 259	90 383 264	2 412 078	81 627 685	261 181	8 755 579
VIII	2 673 259	115 270 389	2 448 428	105 601 430	224 831	9 668 959
IX	2 673 259	161 037 534	2 485 623	149 989 233	187 636	11 048 302
X	2 673 263	363 530 970	2 582 016	351 987 611	91 247	11 543 359
Coeficiente de Gini[2]	0.467					

[1] Los hogares están ordenados en deciles de acuerdo con su ingreso corriente trimestral. Los hogares que tuvieron cero ingreso corriente, se clasifican en el primer decil.

[2] El coeficiente de Gini es una medida de concentración del ingreso: toma valores entre cero y uno. Cuando el valor se acerca a uno, indica que hay mayor concentración del ingreso, en cambio, cuando el valor del Gini se acerca a cero, la concentración del ingreso es menor.

Nota: Los datos son expresados en miles de pesos, motivo por el cual se puede encontrar una diferencia en las cifras totales por cuestiones de redondeo.

Fuente: INEGI, *Nueva Construcción de Variables de la Encuesta Nacional de Ingresos y Gastos de los Hogares* 2008, disponible en http://www.inegi.org.mx/Sistemas/TabuladosBasicos2/TabDirecto.aspx?s=est&c=27298, consultada el 29 de marzo de 2011.

3.1.9 Relación entre la media, la mediana y la moda

Cuando una distribución de frecuencias es simétrica, la media, la mediana y la moda son iguales: $\overline{X} = Med = Mod$. Si la distribución está sesgada a la izquierda, la moda es menor que la mediana y ésta, a su vez, es menor que la media: $Mod < Med < \overline{X}$. En una distribución sesgada a la derecha la relación se invierte, la moda es mayor que la mediana, y ésta a su vez mayor que la media $Mod > Med > \overline{X}$.

Estas relaciones sirven para determinar si una distribución es simétrica y, si no lo es, en qué sentido está sesgada. Se revisa en seguida un ejemplo.

EJEMPLO 3.19

En la tabla siguiente se muestran las distancias, en kilómetros, que recorren 25 estudiantes para ir de sus casas a la escuela. Diga si la distribución es simétrica y, si no lo es, de qué lado está sesgada.

Km recorridos	*f*	*f* acumulada	*Pm*	*f* (*Pm*)
1 a menos de 3	4	4	2	8
3 a menos de 5	9	13	4	36
5 a menos de 7	6	19	6	36
7 a menos de 9	5	24	8	40
9 a menos de 11	1	25	10	10
Suma	25			130

La media aritmética es: $\bar{X} = \dfrac{\sum f_i Pm_i}{\sum f_i} = \dfrac{130}{25} = 5.2$

La posición de la mediana = (25 + 1)/2 = 13, por lo que la clase mediana es la que va de 3 a menos de 5 y, entonces,

$$Med = L_{\text{icmed}} + i\left(\frac{\frac{n+1}{2} - f_{aacmed}}{f_{cmed}}\right) = 3 + 2\left(\frac{13-4}{9}\right)$$

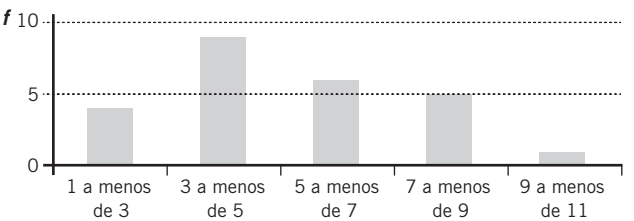
$$= 3 + 2(1) = 3 + 2 = 5$$

La moda es:

$$Mod = L_{\text{icmod}} + i\left(\frac{d_1}{d_1 + d_2}\right) = 3 + 2\left(\frac{5}{5+3}\right)$$

$$= 3 + 2\left(\frac{5}{8}\right) = 3 + 2(0.625) = 3 + 1.25 = 4.25$$

Como $Mod = 4.25 < Med = 5 < \bar{X} = 5.2$, se sabe que la distribución está sesgada a la izquierda, lo cual puede apreciarse visualmente en la gráfica siguiente.



ejercicios 3.1 Medidas de posición o de tendencia central

Media aritmética, mediana, moda y cuartiles

1. En la tabla siguiente se anotan las tasas de interés promedio mensual para los Certificados de la Tesorería de la Federación (Cetes) a 28 días, de marzo de 2009 a febrero de 2011. Determine:

a) La media.
b) La mediana.
c) La moda.
d) Los cuartiles.

Mar. 2009	7.03	Mar. 2010	4.45
Abr. 2009	6.05	Abr. 2010	4.44
Mayo 2009	5.29	Mayo 2010	4.52
Jun. 2009	4.98	Jun. 2010	4.59
Jul. 2009	4.59	Jul. 2010	4.60
Ago. 2009	4.49	Ago. 2010	4.52
Sep. 2009	4.48	Sep. 2010	4.43
Oct. 2009	4.51	Oct. 2010	4.03
Nov. 2009	4.51	Nov. 2010	3.97
Dic. 2009	4.50	Dic. 2010	4.30
Ene. 2010	4.49	Ene. 2011	4.14
Feb. 2010	4.49	Feb. 2011	4.04

Fuente: Banco de México, Tasas y precios de referencia, disponible en: http://www.banxico.org.mx/politica-monetaria-e-inflacion/estadisticas/otros-indicadores/tasas-interes-precios.html, consultado el 29 de marzo de 2011.

2. En la tabla siguiente se presenta la cantidad de minutos que 30 estudiantes invierten para trasladarse de su casa a la escuela. Encuentre:

a) El promedio del tiempo de traslado.
b) La mediana.
c) La moda.
d) Los cuartiles.

29	20	19	15	43
32	21	31	41	25
25	36	43	23	18
23	15	25	33	19
42	16	15	24	32
28	33	17	32	28

3. En el siguiente cuadro se muestran las temperaturas (en grados centígrados) máxima y mínima para diversas ciudades mexicanas cierto día de febrero. Determine:

a) La media aritmética.
b) La mediana.
c) La moda.
d) Los cuartiles de las temperaturas mínimas y de las máximas.

Ciudad	Máx.	Mín.
Acapulco	31	23
Aguascalientes	26	14
Campeche	29	22
Cancún	31	24
Chihuahua	32	17
Ciudad Juárez	34	17
Cuernavaca	30	17
Distrito Federal	23	11
Durango	26	11

(*continúa*)

(*continuación*)

Ciudad	Máx.	Mín.
Guadalajara	27	17
Hermosillo	40	23
La Paz	33	23
Matamoros	34	18
Mazatlán	32	24
Mérida	31	23
Monterrey	31	19
Nuevo Laredo	38	18
Oaxaca	29	18
Puebla	26	14
Puerto Vallarta	32	23
Saltillo	24	10
San Luis Potosí	25	16
Tampico	31	22
Tijuana	29	19
Torreón	31	17
Veracruz	31	22
Villahermosa	31	22

4. En la siguiente tabla se muestra la superficie (en km^2) de los estados y el Distrito Federal de la República Mexicana. Determine:

a) El promedio.
b) La mediana.
c) La moda.
d) Los cuartiles de la superficie por entidad federativa.

Entidad	Superficie
Aguascalientes	5 471
Baja California Norte	69 921
Baja California Sur	73 476
Campeche	50 812
Coahuila	149 982
Colima	5 191
Chiapas	74 211
Chihuahua	244 938
Distrito Federal	1 479
Durango	123 181
Guanajuato	30 491
Guerrero	64 281
Hidalgo	20 813
Jalisco	80 836
México	21 355
Michoacán	59 928
Morelos	4 950
Nayarit	26 979
Nuevo León	64 924
Oaxaca	93 952
Puebla	33 902
Querétaro	11 449
San Luis Potosí	63 068
Sinaloa	58 328
Sonora	182 052
Quintana Roo	50 512
Tabasco	25 267
Tamaulipas	79 384
Tlaxcala	4 016
Veracruz	71 699
Yucatán	38 402
Zacatecas	73 252

5. Se aplica un examen de aptitud a 50 aspirantes al puesto de gerente. Se muestran a continuación las calificaciones finales. Calcule:

a) La media aritmética.
b) La mediana.
c) La moda.
d) Los cuartiles.

25	43	47	59	67	74	78	85	89	94
32	43	53	63	67	75	79	86	89	94
36	43	54	65	67	75	79	86	90	97
39	45	56	65	68	77	85	87	91	97
42	46	57	66	69	77	85	87	93	97

6. Con los datos del ejercicio 2, del tiempo que tardan 30 estudiantes en trasladarse de su casa a su escuela, realice las siguientes tareas:

a) Agrupe la información en una serie de datos y frecuencias.
b) Calcule la media aritmética.
c) Obtenga la mediana.
d) Determine la moda.
e) Calcule los cuartiles.

7. Se prueban 400 focos para determinar su vida útil. El resultado se muestra a continuación. Calcule:

a) La media aritmética.
b) La mediana.
c) La moda y los cuartiles.

Vida útil (días)	Núm. de focos
73	11
75	14
77	16
80	19
82	23
85	35
89	41
91	44
94	47
97	39
99	32
100	28
103	21
106	17
107	13
Total	400

8. En la siguiente tabla se registra el número de clientes que compra determinado número de artículos en un supermercado. Calcule:

a) La media aritmética.
b) La mediana.
c) La moda.
d) Los cuartiles.

Núm. de artículos	Núm. de clientes
1	19
2	15
3	22
4	19
5	24
6	24
7	17
8	9
9	13
10	26
11	32
12	18
13	11
14	24
15	10
16	27
17	14
18	23
19	17

Núm. de artículos	Núm. de clientes
20	30
21	12
22	21
23	18
24	20
25	29
26	12
27	26
28	10
29	16
30	8
31	9
32	11
33	7
34	4
35	3
Total	600

9. Se preguntó a 300 personas el número de libros que leyeron durante el último año. Los resultados se muestran a continuación. Calcule:

a) La media aritmética.
b) La mediana.

Núm. de libros	*f*
0	30
1	39
2	48
3	46
4	42
5	31
6	25
7	11
8	7
9	9
10	6
11	4
12	2
Total	300

10. Se preguntó a 45 estudiantes el número de días a la semana que practican algún deporte, los resultados se muestran en la siguiente tabla. Calcule:

a) La media aritmética.
b) La mediana.
c) La moda.
d) Los cuartiles.

Núm. de días	f
0	6
1	11
2	6
3	13
4	9
Total	45

11. Convierta los datos de traslados de los 30 estudiantes (ejercicios 2 y 6) en una serie de clases y frecuencias y calcule, a partir de esta serie:

a) La media aritmética.
b) La mediana.
c) La moda.
d) Los cuartiles.

12. En la siguiente tabla se muestran los resultados de una encuesta que se hizo a 100 personas acerca del número de horas por semana que utilizan internet. Calcule:

a) La media.
b) La mediana.
c) La moda.
d) Los cuartiles.

x	f
0 a menos de 5	8
5 a menos de 10	23
10 a menos de 15	38
15 a menos de 20	20
20 a menos de 25	11

13. Se preguntó a 150 familias cuánto dinero gastan (en pesos) a la semana en comida rápida, con los resultados que se muestran en la siguiente tabla. Calcule:

a) La media aritmética.
b) La mediana.
c) La moda.
d) Los cuartiles.

x	f
80 a 100	9
101 a 121	24
122 a 142	36
143 a 163	48
164 a 184	21
185 a 205	12
Total	150

14. Se contó durante 45 días el número de visitas a una página web en determinada hora, con los resultados que se muestran en seguida. Calcule:

a) La media aritmética.
b) La mediana.
c) La moda.
d) Los cuartiles.

x	f
80 a menos de 90	3
90 a menos de 100	6
100 a menos de 110	5
110 a menos de 120	10
120 a menos de 130	8
130 a menos de 140	7
140 a menos de 150	4
150 a menos de 160	2
Total	45

15. En una fábrica ensambladora de juguetes se tomó el tiempo en minutos que tardan en armar un carrito 60 trabajadores. En la tabla que aparece en seguida se muestran los resultados. Encuentre:

a) La media aritmética.
b) La mediana.
c) La moda.
d) Los cuartiles.

x	f
8 a 10	8
11 a 13	16
13 a 15	16
15 a 17	14
17 a 19	6
Total	60

Media ponderada

16. En marzo, un inversionista compró 150 acciones de Roma, S.A. de C.V. en $250 cada una. En mayo compró 500 más a $300 cada una. En diciembre compra 200 acciones más a $215. Calcule el precio promedio ponderado por acción.

17. Una fábrica de electrodomésticos fabrica 10 productos. El margen de utilidad del último año fiscal, así como las ventas de cada uno se presentan en la siguiente tabla. Calcule el margen de utilidad medio ponderado.

Producto	Margen de utilidad *M*	Ventas *V*
Abrelatas eléctrico	2.1%	50 000
Batidora	4.3%	26 000
Cuchillo eléctrico	3.6%	33 000

Producto	Margen de utilidad *M*	Ventas *V*
Extractor de jugos	9.4%	120 000
Horno de microondas	12.3%	560 000
Horno eléctrico	8.2%	247 000
Licuadora	7.4%	783 000
Sandwichera	4.2%	67 000
Tostadora	5.5%	96 000
Wafflera	6.1%	43 000
Suma	63.1%	2 025 000

18. En una pastelería se venden 4 tamaños de pastel, familiar, grande, mediano y chico, en 230, 200, 170, 135 y 100, respectivamente. En la siguiente tabla se muestran las ventas del último mes. Calcule el precio promedio.

Tipo de pastel	Precio *P*	Núm. de pasteles *N*
Familiar	200	12
Grande	170	17
Mediano	135	24
Chico	100	13
Total		66

19. En un despacho de asesoría administrativa se cobran $120 por hora para realizar una investigación, $95 por hora por asesoría y $200 por hora por redactar un documento. La semana pasada uno de los colaboradores pasó 8 horas dando asesoría a un cliente, 13 realizando una investigación y 10 redactando un documento. Calcule el cargo medio por sus servicios.

20. En una obra en construcción se paga por hora a los albañiles de la siguiente manera: 27 a $10.50; 16 a $ 13.50; 10 a $17.50, y 4 a $20.50. Calcule el pago medio por hora.

Media armónica

21. Una empresa que cuenta con cinco camiones envía semanalmente carga hasta un almacén que está a 300 kilómetros de distancia. Si, para una semana dada, los camiones viajaron a una velocidad promedio de 80, 90, 70, 95 y 85 kilómetros por hora, calcule la velocidad promedio de todos los camiones en esa semana.

22. En una oficina hay 2 ventanillas; una la atiende un empleado muy eficiente y la otra un novato; el empleado eficiente atiende a cada cliente en 5 minutos y el novato se tarda 10. ¿Cuál debe ser la velocidad de 2 empleados que trabajen al mismo ritmo y que atiendan el mismo número de clientes que los 2 empleados disparejos?

23. En una competencia participan 5 pilotos, que realizan un recorrido de 1 000 kilómetros. Se muestra a continuación el tiempo en segundos que tardó cada uno de ellos en terminar el trayecto. Calcule el tiempo medio de la competencia.

Piloto	Tiempo
A	50
B	60
C	100
D	74
E	83

24. Una empresa agrícola dedicada a la producción de plátano tiene 10 terrenos de siembra de distintos tamaños. En la última cosecha se recolectó cierto número de kilogramos de cada uno de ellos. Calcule la producción media de plátanos en los 10 terrenos.

Terreno	Producción
A	1 500
B	2 100
C	1 234
D	2 356
E	3 211
F	1 123
G	820
H	3 452
I	2 678
J	1 320

25. Se tomó el tiempo en minutos que 15 secretarias de una oficina corporativa tardan en mecanografiar un documento de 2 cuartillas. Calcule el tiempo de captura promedio.

Nombre	Tiempo de captura
Ana	6.4
Andrea	4.8
Angélica	3.5
Carmen	4.7
Carolina	5.2
Daniela	5.1
Estela	6.3
Fernanda	4.5
Gabriela	4.7
Guadalupe	3.2
Isabel	3.6
María	3.2
Rebeca	5.2
Rosa	5.5
Susana	4.9

Media geométrica

26. A continuación se presenta el incremento porcentual en las ventas de Vázquez y Asociados, S.A., durante los últimos 5 años. Calcule la media geométrica del incremento porcentual.

9.4	12.7	10.5	12.9	14.2

27. En un experimento químico se tomaron las siguientes temperaturas (en grados centígrados): 13.4, 12.5, 11.8, 13.6. Calcule la temperatura promedio.

28. Las tasas de interés de 10 bonos se muestran a continuación. Calcule la tasa promedio.

5%	2%	3%	7%	4%	4.5%	2.5%	3.5%	6%	4%

29. A continuación se muestra el incremento porcentual del valor de las acciones de la compañía Harinas, S.A., durante los últimos 15 años. Calcule la media geométrica

5.3	7.4	7.6	8.8	9.3
6.2	7.8	7.9	9.2	9.9
6.9	8.1	8.3	9.5	10.1

30. En 2000 había 33 millones de suscriptores a los servicios de televisión satelital. Para 2005 el número de suscriptores aumentó a 61 millones. Calcule el incremento promedio anual.

Promedios móviles

31. En la tabla siguiente se muestra el tipo de cambio para solventar obligaciones en moneda extranjera, con valores a fin de mes, de marzo de 2009 a febrero de 2011, según datos del Banco de México.

a) Calcule los promedios móviles de 3 y de 5 meses.
b) Grafique los promedios móviles que calculó en el inciso *a*).
c) Comente las diferencias que observa entre ambos promedios móviles.

Tipo de cambio: pesos por dólar estadounidense para solventar obligaciones denominadas en moneda extranjera, fecha de determinación (FIX), cotizaciones promedio	
Mar. 2009	14.6695
Abr. 2009	13.4367
Mayo 2009	13.1621
Jun. 2009	13.3418
Jul. 2009	13.3654
Ago. 2009	13.0080
Sep. 2009	13.4212
Oct. 2009	13.2257
Nov. 2009	13.1094
Dic. 2009	12.8631
Ene. 2010	12.8019
Feb. 2010	12.9424
Mar. 2010	12.5737
Abr. 2010	12.2302
Mayo 2010	12.7428
Jun. 2010	12.7193
Jul. 2010	12.8189
Ago. 2010	12.7695
Sep. 2010	12.7997
Oct. 2010	12.4374
Nov. 2010	12.3391
Dic. 2010	12.3885
Ene. 2011	12.1258
Feb. 2011	12.0703

Fuente: Banco de México, disponible en: http://www.banxico.gob.mx/sistema-financiero/estadisticas/mercado-cambiario/tipos-cambio.html, consultado el 29 de marzo de 2011.

32. En la tabla siguiente se muestran los datos de billetes y monedas en poder del público, de marzo de 2009 a febrero de 2011, según datos del Banco de México.

a) Calcule los promedios móviles de 4 y de 8 meses.
b) Grafique los promedios móviles que calculó en el inciso *a*).
c) Comente las diferencias que observa entre ambos promedios móviles.

Mar. 2009	533 819 278
Abr. 2009	537 331 410
Mayo 2009	536 606 477
Jun. 2009	531 629 602
Jul. 2009	541 061 603
Ago. 2009	529 495 386
Sep. 2009	524 785 931
Oct. 2009	534 520 513
Nov. 2009	547 652 896
Dic. 2009	631 937 880
Ene. 2010	600 421 918
Feb. 2010	584 170 708

Mar. 2010	597 193 947
Abr. 2010	574 362 947
Mayo 2010	582 623 188
Jun. 2010	577 815 451
Jul. 2010	593 182 402
Ago. 2010	584 119 067
Sep. 2010	588 091 817
Oct. 2010	590 029 643
Nov. 2010	605 690 050
Dic. 2010	693 423 114
Ene. 2011	648 030 446
Feb. 2011	638 281 570

Fuente: Banco de México, disponible en: http://www.banxico.org.mx/billetes-y-monedas/estadisticas/billetes-y-monedas-en-circulacion/pagina-billetes-monedas-en-ci.html, consultado el 29 de marzo de 2011.

33. En la tabla siguiente se muestran los índices de morosidad de la banca múltiple en México, de enero de 2009 a diciembre de 2010, con datos de la Comisión Nacional Bancaria y de Valores.

a) Calcule los promedios móviles de 3 y de 4 meses.
b) Grafique los promedios móviles que obtuvo en el inciso *a*).
c) Comente las diferencias que observa entre ambos promedios móviles.

Fecha	Índice de morosidad de la banca múltiple
Ene. 2009	3.36
Feb. 2009	3.42
Mar. 2009	3.43
Abr. 2009	3.63
Mayo 2009	3.86
Jun. 2009	3.84
Jul. 2009	3.76
Ago. 2009	3.61
Sep. 2009	3.43
Oct. 2009	3.38
Nov. 2009	3.34
Dic. 2009	3.08
Ene. 2010	3.06
Feb. 2010	2.95
Mar. 2010	2.77
Abr. 2010	2.86
Mayo 2010	2.88
Jun. 2010	2.68
Jul. 2010	2.71
Ago. 2010	2.61
Sep. 2010	2.49
Oct. 2010	2.46
Nov. 2010	2.47
Dic. 2010	2.33

Fuente: Comisión Nacional Bancaria y de Valores, *Información estadística detallada*, disponible en: http://portafoliodeinformacion.cnbv.gob.mx/BM/Paginas/capital.aspx, consultado el 29 de marzo de 2011.

3.2 Medidas de dispersión

En este apartado se revisan las medidas estadísticas que se usan para analizar qué tan dispersos o separados están los datos entre sí. Una de estas medidas es el **rango**, el cual se calcula simplemente como la diferencia entre el mayor y el menor de los valores o, en otras palabras, es igual al máximo valor menos el menor y, como puede apreciarse, mide qué tan separados están los datos en esta sencilla forma. Por su parte, como se verá, las que son las 2 principales medidas de dispersión, la **desviación estándar** y la **varianza**, miden la dispersión de los datos alrededor de la media aritmética. Para calcular estas medidas debe hacerse primero la diferencia entre cada dato de la serie y su media que es como se mide, básicamente, la dispersión.

Rango. Diferencia entre el mayor y el menor de los valores; mide qué tan separados están los datos.
Desviación estándar. Mide la dispersión de los datos alrededor de la media aritmética.
Varianza. Mide la dispersión de los datos alrededor de la media aritmética.

Con los ejemplos se comprenderán bien los detalles.

Las medidas de dispersión que se revisan en esta sección son: rango, desviación media, desviación intercuartílica, varianza y desviación estándar. Se incluye una sección sobre aplicaciones de la desviación estándar que ilustra su uso con el coeficiente de variación, el teorema de Chebyshev y la muy conocida y útil distribución normal.

3.2.1 Rango

Tal como se mencionó, el rango es la diferencia entre los valores mayor y menor.

EJEMPLO 3.20

Se reproduce en seguida como tabla 3.17 la tabla 3.11 que contiene los datos de edades de padres de familia agrupados en una serie de datos y frecuencias.

Tabla 3.17 Edades de 60 padres de familia agrupados en datos y frecuencias

Edad x	Frecuencia f
46	1
48	1
49	2
50	4
52	5
53	3
54	7
55	4
56	4
57	1
58	5
63	5
64	2
65	7
68	3
70	1
75	1
78	2
Sumas	60

Como los valores máximo y mínimo son 78 y 46, respectivamente, es fácil ver que el rango de estos datos es $78 - 46 = 32$ y es el mismo valor que se hubiera encontrado si se hubiera partido de la serie simple, ya que en el listado original de las 60 edades, los valores máximo y mínimo son los mismos.

Sin embargo, y al igual que ha sucedido antes, cuando se determina el rango a partir de una serie de clases y frecuencias se pueden dar diferencias, en este caso por la forma en la que se agrupan los datos en clases.

EJEMPLO 3.21

Se reproduce en seguida como tabla 3.18 la serie de clases y frecuencias en la que se agruparon los datos de las edades.

Tabla 3.18 Las edades de 60 padres de familia agrupadas en una serie de clases y frecuencias

x	f
45 a menos de 50	4
50 a menos de 55	21
55 a menos de 60	14
60 a menos de 65	7
65 a menos de 70	10
70 a menos de 75	1
75 a menos de 80	3
Totales	60

Aquí el valor menor es el límite inferior de la primera clase, 45, y el mayor valor es 80, el límite superior de la última clase, por lo que el rango es de 35 y es diferente al que se encontró antes, por las razones mencionadas.

NOTA En lo sucesivo, no se indica que no existen funciones específicas de Excel para datos agrupados; simplemente se ignoran estos casos para Excel.

EXCEL Así como no existen funciones de Excel para calcular las medidas de posición, tampoco las hay para obtener las medidas de dispersión en datos agrupados, por lo que sólo existen las que se aplican a series simples y tampoco hay una para el rango, aunque es fácil ver que se puede determinar mediante una resta del máximo y

el mínimo. Por ejemplo, si se tiene un conjunto de datos en las celdas A1:E5, la siguiente diferencia produce su rango:

=Max(A1:E5)-MIN(A1:E5)

3.2.2 Desviación media

Se mencionó que una base para medir dispersión que es importante en estadística es la que parte de las diferencias entre cada dato y su media, $X_i - \overline{X}$. La **desviación media** es el promedio de los valores absolutos de estas diferencias. En el ejemplo siguiente se ilustra la razón por la cual se utilizan estos valores absolutos y no las diferencias mismas.

Desviación media. Promedio de los valores absolutos de las diferencias entre cada dato y su media.

Ejemplo 3.22

En la tabla 3.19 se muestran las distancias, en kilómetros, que recorren 8 estudiantes para ir de su casa a la universidad.

Tabla 3.19 Datos para el ejemplo 3.22

Estudiante	Km recorridos (variable X)	$X_i - \overline{X}$	$\lvert X_i - \overline{X} \rvert$
A	2	- 4	4
B	8	2	2
C	12	6	6
D	6	0	0
E	10	4	4
F	4	–2	2
G	3	–3	3
H	3	–3	3
Suma	48	0	24

Como el promedio de kilómetros recorridos es: $\overline{X} = \frac{48}{8} = 6$, la columna de $X_i - \overline{X}$ se calculó restando a cada dato esta media. Como puede verse en la tabla, la suma de estas diferencias es cero y ésta es la razón por la que no puede utilizarse esta suma para medir dispersión; esta suma de diferencias entre cada dato y su media siempre es de 0. Esto es también la razón por la cual se utilizan los valores absolutos de estas diferencias para medir la desviación media, que es:

$$DM = \frac{\sum \left| X_i - \overline{X} \right|}{n} \tag{3.13}$$

Y, con los datos del ejemplo,

$$DM = \frac{\sum \left| X_i - \overline{X} \right|}{n} = \frac{24}{8} = 3$$

Excel La función de Excel que calcula la desviación media es DESVPROM. En el caso anterior, la función podría ser =DESVPROM(A1:H1), la cual arrojaría el mismo valor de 3.

Ejemplo 3.23

La desviación media de una serie de datos y frecuencias

Calcular la desviación media de los datos de edades agrupados en serie de datos y frecuencias.

Solución:

En la tabla 3.20 se reproduce la serie correspondiente, junto con las operaciones necesarias.

Tabla 3.20 Edades de 60 padres de familia agrupados en datos y frecuencias

x (edad)	f	$f(x)$	$X_i - \overline{X}$	$f(X_i - \overline{X})$	$\lvert f(X_i - \overline{X}) \rvert$
46	1	46	–12.117	–12.117	12.117
48	1	48	–10.117	–10.117	10.117
49	2	98	–9.117	–18.234	18.234

(*continúa*)

Tabla 3.20 (*continuación*)

x (edad)	f	f(x)	$X_i - \bar{X}$	$f(X_i - \bar{X})$	$\lvert f(X_i - \bar{X})\rvert$
50	4	200	–8.117	–32.468	32.468
52	7	364	–6.117	–42.819	42.819
53	3	159	–5.117	–15.351	15.351
54	7	378	–4.117	–28.819	28.819
55	4	220	–3.117	–12.468	12.468
56	4	224	–2.117	–8.468	8.468
57	1	57	–1.117	–1.117	1.117
58	5	290	–0.117	–0.585	0.585
63	5	315	4.883	24.415	24.415
64	2	128	5.883	11.766	11.766
65	7	455	6.883	48.181	48.181
68	3	204	9.883	29.649	29.649
70	1	70	11.883	11.883	11.883
75	1	75	16.883	16.883	16.883
78	2	156	19.883	39.766	39.766
	60	3 487		–0.02	365.106

Con los datos de la tercera columna se calcula la media:

$$\bar{X} = \frac{\sum f X_i}{n} = \frac{3\,487}{60} = 58.117$$

Con este valor de la media se hacen las operaciones marcadas para las columnas 3, 4 y 5, y del total de esta última columna se obtiene la desviación media:

$$DM = \frac{\sum \left| f(X - \bar{X}) \right|}{\sum f} \qquad \textbf{(3.14)}$$

Con los datos del ejemplo:

$$DM = \frac{\sum \left| f\left(X - \bar{X}\right) \right|}{\sum f} = \frac{365.106}{60} = 6.085$$

EJEMPLO 3.24

Desviación media en una serie de clases y frecuencias

En la tabla 3.21 se muestran los datos de las edades, agrupadas en una serie de clases y frecuencias, y se calcula con esos datos la desviación media.

Tabla 3.21 Las edades de 60 padres de familia agrupadas en una serie de clases y frecuencias

x	f
45 a menos de 50	4
50 a menos de 55	21
55 a menos de 60	14
60 a menos de 65	7
65 a menos de 70	10
70 a menos de 75	1
75 a menos de 80	3
Totales	60

Los cálculos para la desviación media se muestran en la siguiente tabla:

x	f	Pm	fPm	$Pm - \bar{X}$	$f(Pm - \bar{X})$	$\lvert f(Pm - \bar{X})\rvert$
45 a menos de 50	4	47.5	190	–9.08	–36.32	36.32
50 a menos de 55	21	52.5	1102.5	–4.08	–85.68	85.68
55 a menos de 60	14	57.5	805	0.92	12.88	12.88
60 a menos de 65	7	62.5	437.5	5.92	41.44	41.44
65 a menos de 70	10	67.5	675	10.92	109.2	109.2
70 a menos de 75	1	72.5	72.5	15.92	15.92	15.92
75 a menos de 80	3	77.5	232.5	20.92	62.76	62.76
Totales	60		3 515			364.2

Como puede verse en la tabla anterior, la media aritmética es,

$$\bar{X} = \frac{3\,515}{60} = 58.58$$

La desviación media, o desviación promedio es,

$$DM = \frac{\sum \left| f(Pm - \bar{X}) \right|}{\sum f} \qquad \textbf{(3.15)}$$

Con los datos del ejemplo:

$$DM = \frac{364.2}{60} = 6.07$$

3.2.3 Desviación intercuartílica

Desviación intercuartílica. Diferencia entre el tercer y el primer cuartiles.

La **desviación intercuartílica** es simplemente la diferencia entre el tercer y el primer cuartiles. En símbolos:

$$DI = Q_3 - Q_1 \tag{3.16}$$

Ejemplo 3.25

En el ejemplo 3.18, que se refería al número de pasajeros que viajaron en 100 vuelos de Aerolíneas Mayas, se encontraron los siguientes valores para los cuartiles 1 y 3: $Q_1 - 71.15$ y $Q_1 - 92.083$, por lo que:

$$DI = Q_3 - Q_1 = 92.083 - 71.15 = 20.933$$

3.2.4 Varianza y desviación estándar

Desviación estándar. Raíz cuadrada de la varianza.

La varianza junto con la **desviación estándar** que, como se verá, es la raíz cuadrada de aquélla, son las 2 principales medidas de dispersión y se utilizan con mucha frecuencia en diversos procedimientos estadísticos.

Siendo, entonces, medidas tan importantes, es conveniente conocerlas con detenimiento y, para ello, vale la pena enfatizar desde ahora que estas 2 medidas miden dispersión respecto a la media aritmética, ya que la varianza es el promedio de los cuadrados de las desviaciones de cada dato en relación con su media o, en símbolos:

$$S^2 = \frac{\sum (X_i - \bar{X})^2}{n} \tag{3.17}$$

La fórmula anterior resume el procedimiento para calcular la varianza de una muestra; de ahí el símbolo S^2. La fórmula que describe el mismo procedimiento cuando se calcula la varianza de la población es:

$$\sigma^2 = \frac{\sum (X_i - \mu)^2}{n} \tag{3.18}$$

Abundando en la descripción de esta fórmula, para calcular la varianza, en primer lugar, se resta la media de cada dato [éstas son las diferencias $(X_i - \mu)$]. Después se elevan al cuadrado estas diferencias $(X_i - \mu)^2$, para posteriormente sumarlas $\sum(X_i - \mu)^2$. Con esto se tiene la suma de los cuadrados de las diferencias y si esto se divide entre su número, n, se tiene ya el promedio de los cuadrados de las desviaciones de cada dato respecto a su media, o sea la varianza.

Varianza. Promedio de cuadrados.

Puesto que la **varianza** es un promedio de cuadrados, se calcula también la raíz cuadrada de esta varianza, con lo que se tiene una medida en las unidades originales, ya no cuadrados. Dicha raíz cuadrada de la varianza es, precisamente, la desviación estándar. En símbolos:

$$S = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n}} \tag{3.19}$$

Y, para una población:

$$\sigma = \sqrt{\frac{\sum (X_i - \mu)^2}{n}} \tag{3.20}$$

Ejemplo 3.26

Una fábrica de dulces elabora 10 diferentes productos. A continuación se presentan los costos de producción por cada 100 piezas de las diferentes golosinas y se calculan su varianza y su desviación estándar.

Solución:

Producto	Costo	$X_i - \bar{X}$	$(X_i - \bar{X})^2$
Bombón	$20	–16.5	272.25
Caramelo	$33	–3.5	12.25
Caramelo suave	$41	4.5	20.25
Chocolate	$63	26.5	702.25
Cocada	$38	1.5	2.25
Mentas	$27	–9.5	90.25
Nuez garapiñada	$56	19.5	380.25
Oblea	$19	–17.5	306.25
Paleta	$16	–20.5	420.25
Tamarindo	$52	15.5	240.25
Total	365	0	2246.5

$$\bar{X} = \frac{\sum X_i}{n} = \frac{365}{10} = 36.5$$

$$S^2 = \frac{\sum (X_i - \bar{X})^2}{n} = \frac{2\,446.5}{10} = 244.65$$

$$S = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n}} = \sqrt{\frac{2\,446.5}{10}} = 15.64$$

Excel Las funciones que calculan la varianza y la desviación estándar son VARP y DESVESTP. Así, se podrían calcular estas medidas para los valores anteriores mediante =VARP(B2:B11), y =DESVESTP(B2:B11), que arrojan los mismos resultados calculados antes.

Recuérdese que el rango de celdas anotado en esas funciones (B2:B11) puede variar según la posición en que se coloquen los datos en la hoja de Excel.

El ejemplo anterior ilustra el procedimiento para calcular la varianza y la desviación estándar para una serie simple. En los 2 ejemplos siguientes se ilustran los procedimientos para calcular estas medidas para datos agrupados: para una serie de datos y frecuencias y para una serie de clases y frecuencias.

Ejemplo 3.27

La varianza y la desviación estándar en una serie de datos y frecuencias

Se realizó un estudio para observar cuántas personas se formaron a cierta hora en los cajeros automáticos del banco MexBanc localizados en 25 establecimientos comerciales. En la tabla siguiente se muestran los resultados; calcule la varianza y la desviación estándar.

Solución:

Personas en fila	f	$f(X_i)$	$X_i - \bar{X}$	$(X_i - \bar{X})^2$	$f(X_i - \bar{X})^2$
0	3	0	–2.32	5.3824	16.1472
1	4	4	–1.32	1.7424	6.9696
2	8	16	–0.32	0.1024	0.8192
3	6	18	0.68	0.4624	2.7744
4	1	4	1.68	2.8224	2.8224
5	2	10	2.68	7.1824	14.3648
6	1	6	3.68	13.5424	13.5424
Total	25	58			57.44

Aquí conviene recordar que en los datos agrupados con frecuencias, cada frecuencia indica el número de veces que se repite cada dato. Por ejemplo, el 0 asociado a una frecuencia de 3 dice que en 3 de los cajeros automáticos analizados no había nadie esperando en la fila y, por ello, la necesidad de multiplicar cada dato por su frecuencia para, una vez sumados estos productos, dividir entre la suma de las frecuencias para encontrar la media aritmética.

Se desea enfatizar esto pues es un error común en estudiantes que por primera vez revisan este tema. Así, para calcular aquí la varianza y la desviación estándar, una vez que se calculó la media, se resta ésta de cada dato (número de personas en la fila), $X_i - \bar{X}$ para, después, elevar estas diferencias al cuadrado $(X_i - \bar{X})^2$ y, entonces, multiplicar estos cuadrados de diferencias por la frecuencia correspondiente $f(X_i - \bar{X})^2$. El orden de las columnas es, precisamente, el orden en el que deben llevarse a cabo las operaciones. Resumiendo este procedimiento en una fórmula se tiene que la varianza para una serie de datos y frecuencias es

$$S^2 = \frac{\sum f(X_i - \bar{X})^2}{\sum f_i} \qquad \textbf{(3.21)}$$

Si se calculara para una población, los símbolos serían:

$$\sigma^2 = \frac{\sum f(X_i - \mu)^2}{N} \qquad \textbf{(3.22)}$$

Entonces, para los datos del ejemplo:

$$\bar{X} = \frac{\sum f_i X_i}{\sum f_i} = \frac{58}{25} = 2.32$$

$$S^2 = \frac{\sum f\left(X_i - \bar{X}\right)^2}{\sum f_i} = \frac{57.44}{25} = 2.3 \qquad \text{y} \qquad S = \sqrt{\frac{\sum f\left(X_i - \bar{X}\right)^2}{\sum f_i}} = \sqrt{\frac{57.44}{25}} = 1.52$$

EJEMPLO 3.28

La varianza y la desviación estándar en una serie de clases y frecuencias

Se realizó un estudio para determinar cuánto gastaron en publicidad 65 de las más importantes empresas establecidas en México, con los resultados que se muestran en la siguiente tabla. Calcule la varianza, la desviación estándar y el coeficiente de variación.

Gastos publicitarios (millones de pesos)	*f*
10 a menos de 25	4
25 a menos de 40	7
40 a menos de 55	9
55 a menos de 70	12
70 a menos de 80	15
80 a menos de 95	8
95 a menos de 110	5
110 a menos de 125	3
125 a menos de 140	2
Total	65

Solución:

Al igual que se hizo antes con la serie de datos y frecuencias, se debe enfatizar la importancia de no olvidar el significado de las frecuencias en series de este tipo. Para calcular la media, en primer lugar se encuentra el punto medio de clase, que se utiliza para representar a todos los elementos de cada clase y sabemos cuántos son, precisamente, a través de la frecuencia. Después se multiplica este punto medio por la frecuencia para, al dividir la suma de estos productos por la suma de las frecuencias (que arroja el número total de elementos, *n*), obtener la media aritmética.

En la tabla inferior se resumen los cálculos: 1) se resta la media de cada punto medio, 2) se elevan estas diferencias al cuadrado, 3) se multiplican estos cuadrados por la frecuencia y la varianza es el cociente entre la suma de estos productos y la suma de las frecuencias. Los símbolos para obtener una varianza calculada a partir de datos muestrales son:

$$S^2 = \frac{\sum f_i\left(Pm_i - \bar{X}_i\right)^2}{\sum f_i} \qquad \textbf{(3.23)}$$

Los símbolos para la varianza calculada a partir de una población son:

$$\sigma^2 = \frac{\sum f(Pm_i - \mu)^2}{N} \qquad \textbf{(3.24)}$$

Gastos publicitarios (millones de pesos)	*f*	*Pm*	*fPm*	$Pm - \bar{X}$	$(Pm - \bar{X})^2$	$f(Pm - \bar{X})$	$f(Pm - \bar{X})^2$
10 a menos de 25	4	17.5	70	–52.615	2 768.34	–210.46	1 1073.36
25 a menos de 40	7	32.5	227.5	–37.615	1 414.89	–263.305	9 904.23
40 a menos de 55	9	47.5	427.5	–22.615	511.44	–203.535	4 602.96
55 a menos de 70	12	62.5	750	–7.615	57.99	–91.38	695.88
70 a menos de 80	15	77.5	1162.5	7.385	54.54	110.775	818.1
80 a menos de 95	8	92.5	740	22.385	501.09	179.08	4 008.72
95 a menos de 110	5	107.5	537.5	37.385	1 397.64	186.925	6 988.2
110 a menos de 125	3	122.5	367.5	52.385	2 744.19	157.155	8 232.57
125 a menos de 140	2	137.5	275	67.385	4 540.74	134.77	9 081.48
Total	65		4 557.5			0.025	5 5405.5

$$\bar{X} = \frac{\sum f_i Pm_i}{\sum f_i} = \frac{4\,557.5}{65} = 70.115$$

$$S^2 = \frac{\sum f_i\left(Pm - \bar{X}\right)^2}{\sum f_i} = \frac{55\,405.5}{65} = 852.39$$

$$S = \sqrt{\frac{\sum f_i\left(Pm_i - \bar{X}\right)}{\sum f_i}} = \sqrt{\frac{55\,405.5}{65}} = 29.19$$

3.2.4.1 Uso de la varianza y la desviación estándar muestrales como estimadores

Las fórmulas que se anotaron antes para estas 2 medidas de dispersión para una muestra deben modificarse cuando se utilizan en estadística inferencial para estimar el valor correspondiente de una población.

En este tema de la inferencia estadística se utilizan medidas muestrales (como la varianza de una muestra) para estimar el correspondiente valor de la población (la varianza de la población). En este caso, cuando se usa la varianza de la muestra para estimar la de la población, se deben modificar los cálculos de la varianza (y con ellos, los de la desviación estándar). Estas fórmulas modificadas son:

$$S^2 = \frac{\sum (X_i - \bar{X})^2}{n - 1} \tag{3.25}$$

$$S = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n - 1}} \tag{3.26}$$

Como puede verse en las fórmulas anteriores, la diferencia con las utilizadas antes es el numerador, que se convierte en $n - 1$, en vez de ser simplemente, n. La razón de esta modificación se revisa con detalle en el capítulo 7 "Muestreo y distribuciones muestrales", con los temas de distribuciones muestrales y el de estimadores sesgados e insesgados. Por lo pronto baste esta breve mención, que es importante tener en cuenta desde este momento y que, insistiendo, se explica con detalle en ese capítulo 7 y se aplica en los capítulos subsiguientes que tratan de la inferencia estadística.

3.2.5 Aplicaciones comunes de la desviación estándar

Se revisan aquí cuatro maneras en las que se suele aplicar la desviación y que ayudan a comprender su importancia.

3.2.5.1 Coeficiente de variación

Coeficiente de variación. Cociente entre la desviación estándar y la media aritmética multiplicado por 100.

La varianza y la desviación estándar son medidas absolutas, porque se basan en los valores originales de las variables correspondientes. El **coeficiente de variación** que es, simplemente, el cociente entre la desviación estándar y la media aritmética multiplicado por 100, es una medida relativa de dispersión ya que esa forma de cálculo implica que su valor indica qué proporción de la media representa la desviación estándar.

El coeficiente de variación en símbolos:

$$CV = \frac{S}{\bar{X}}(100) \tag{3.27}$$

En el ejemplo de la fábrica de golosinas, el 3.26, se encontró que

$$\bar{X} = 36.5$$
$$s = 15.64$$

De donde el coeficiente de variación es

$$CV = \frac{S}{\bar{X}}(100) = \frac{15.64}{36.5}(100) = 42.85$$

Por ello se dice que este coeficiente es una medida de dispersión relativa porque, como puede verse en la forma en que se calcula, indica qué proporción de la media representa la desviación estándar. Ese 42.85 que se obtuvo significa que la desviación estándar representa 42.85% de la media.

En el ejemplo 3.27, que trata del número de personas formadas en una fila a cierta hora, se encontró que:

$$\bar{X} = 2.32$$
$$s = 1.51$$

Por lo que:

$$CV = \frac{S}{\bar{X}}(100) = \frac{1.51}{2.32}(100) = 65.09$$

El cual indica que la desviación estándar representa 65.09% de la media aritmética. También, en el ejemplo 3.28, que trataba de gastos de publicidad se encontró que:

$$\overline{X} = 70.115$$
$$s = 29.19$$

De donde el coeficiente de variación es:

$$CV = \frac{S}{\overline{X}}(100) = \frac{29.19}{70.115}(100) = 41.63$$

Que indica que la desviación estándar representa 41.63% de la media aritmética.

En el ejemplo siguiente se resumen los valores encontrados para estas 3 medidas en los últimos 3 ejemplos para ilustrar cómo es que permiten hacer comparaciones útiles respecto a la dispersión de los datos entre poblaciones diferentes.

Ejemplo 3.29

En la tabla siguiente se resumen los valores de las 3 medidas de dispersión analizadas en los 3 ejemplos anteriores.

Ejemplo / medida	Varianza	Desviación estándar	Coeficiente de variación	Media aritmética
Costos de producción	244.65	15.64	42.85	36.5
Personas formadas	2.3	1.51	65.09	2.32
Gastos de publicidad	852.39	29.19	41.63	70.115

Pudiera existir la inquietud de comparar las varianzas o las desviaciones estándar de estas 3 muestras, pero no sería apropiado porque la desviación estándar de 1.51 para las personas que hacen fila en los cajeros automáticos son valores que varían sólo entre 0 y 6, y por ello tienen una media aritmética de apenas 2.32, en tanto que los datos con mayor desviación estándar, que son los de los gastos de publicidad, tienen una media de 70.115, por lo que es natural que tengan una desviación estándar mayor: la comparación directa entre las desviaciones estándar (o las varianzas) de poblaciones o muestras con medias muy distintas no es apropiada porque son valores absolutos. Sin embargo, la comparación directa entre coeficientes de variación sí lo es porque se trata de valores relativos; como se revisó antes, estos coeficientes de variación dicen qué proporción de la media representa su correspondiente desviación estándar. En estos términos, se puede ver que la muestra que tiene más variación en relación con su media es la de las personas que hacen fila, porque representan el 65.09 de su media y, por su parte, los que tienen menor dispersión respecto a su media son los datos de los gastos de publicidad.

3.2.5.2 Teorema de Chebyshev

Formulado por el matemático ruso P.L. Chebyshev (1821-1894), el **teorema de Chebyshev** permite determinar la proporción mínima de valores que se encuentran en un número específico de desviaciones estándar en relación con la media.

Teorema de Chebyshev. Determina la proporción mínima de valores que se encuentran en un número específico de desviaciones estándar en relación con la media.

El teorema de Chebyshev establece que, para cualquier conjunto de datos, al menos $(1-1/K^2)$ de las observaciones están dentro de K desviaciones estándar de la media, en donde K es cualquier número mayor que 1.

$$1 - \left[\frac{1}{K^2}\right]$$

Si formamos un intervalo desde $K = 3$ desviaciones estándar por encima de la media, hasta 3 desviaciones estándar por debajo de ella, se tiene que

$$1 - \left[\frac{1}{3^2}\right] = 0.8889$$

Lo cual quiere decir que, para cualquier conjunto de datos del total de observaciones, cuando menos 88.89% de ellas se encuentran en el intervalo de la media más 3 desviaciones estándar y la media menos 3 desviaciones estándar.

Ejemplo 3.30

En el ejemplo 3.3, donde se muestran las edades de 60 padres de familia, la media fue de 58.58, se calcula la desviación estándar y tenemos:

X	f	Pm	fPm	$Pm - \bar{X}$	$(Pm - \bar{X})^2$	$f(Pm - \bar{X})^2$
45 a menos de 50	4	47.5	190	–11.08	122.77	491.08
50 a menos de 55	21	52.5	1 102.5	–6.08	36.97	776.37
55 a menos de 60	14	57.5	805	–1.08	1.17	16.38
60 a menos de 65	7	62.5	437.5	3.92	15.37	107.59
65 a menos de 70	10	67.5	675	8.92	79.57	795.7
70 a menos de 75	1	72.5	72.5	13.92	193.77	193.77
75 a menos de 80	3	77.5	232.5	18.92	357.97	1 073.91
Totales	60		3 515			3 454.8

$$S = \sqrt{\frac{\sum f_i\left(Pm_i - \bar{X}\right)^2}{\sum f_i}} = \sqrt{\frac{3\,454.8}{60}} = 7.59$$

Se quiere saber qué proporción de los padres de familia tiene edad dentro del intervalo de *a*) $K = 2$ desviaciones estándar, y de *b*) $K = 3$ desviaciones estándar alrededor de la media, y cuál es el intervalo de cada uno de ellos.

Solución:

a) $K = 2$ desviaciones estándar.

$$1 - \left[\frac{1}{2^2}\right](100) = 75\%$$

2 desviaciones estándar = (7.59)(2) = 15.18
Entonces el intervalo es:

$$58.58 - 15.18 = 43.4$$
$$58.58 + 15.18 = 73.76$$

Es decir, de acuerdo al teorema de Chebyshev, cuando menos 75% de los padres tienen edades entre 43.4 y 73.76 años y revisando la serie de datos, se puede ver que sólo 3 de las 60 personas están fuera de este rango o, en otras palabras, que 57 de 60, o 95% están dentro de ese rango lo cual, a su vez, verifica que se cumple el teorema de Chebyshev.

b) $K = 3$ desviaciones estándar

$$1 - \left[\frac{1}{3^2}\right](100) = 88.89\%$$

Tres desviaciones estándar = (7.59)(3) = 22.77
Entonces el intervalo es:

$$58.58 - 22.77 = 35.81$$
$$58.58 + 22.77 = 81.35$$

Es decir, de acuerdo al teorema de Chebyshev, cuando menos 88.89% de las padres tienen edades entre 35.81 y 81.35 años.

Al igual que en el inciso anterior, si se revisa el conjunto de datos se podrá comprobar que, efectivamente, se cumple la regla.

Distribución normal o **campana de Gauss.** Distribución de probabilidad continua de amplia aplicación estadística y en otras disciplinas.

3.2.5.3 Desviación estándar y distribución normal

La **distribución normal** es una distribución de probabilidad continua de muy amplia aplicación en estadística y en otras disciplinas; es ampliamente conocida su forma de campana, por lo que también se le conoce como *campana de Gauss,* en honor del matemático, astrónomo y físico alemán Carl Friedrich Gauss (1777-1855).

Se revisa a detalle esta distribución en el tema de las distribuciones de probabilidad continuas que se trata en el capítulo 6 "Distribuciones continuas de probabilidad". Aquí bastará con mencionar una aplicación de la desviación estándar relacionada con esta importante distribución.

Se sabe, y tiene numerosas aplicaciones, que el área contenida debajo de la curva, por encima del eje horizontal y entre una desviación estándar a la derecha y otra desviación estándar a la izquierda de la media contiene 68.26% del total del área bajo la curva, tal como se ilustra en la figura 3.4.

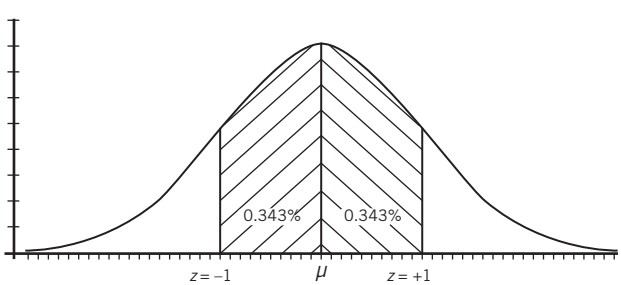


Figura 3.4 Área bajo una curva normal: una desviación estándar alrededor de la media.

Se usa la *z* para representar la desviación estándar de lo que se conoce como la **distribución normal estándar**, definida así porque tiene media aritmética de 0 y desviación estándar de 1. Por ello, en estas condiciones, una *z* arriba y una *z* debajo de la media, significan una desviación estándar por arriba y una desviación estándar por debajo de la media. En esa misma figura se puede apreciar que, como la curva es simétrica, la proporción de área que se encuentra entre la media y una desviación estándar es de 0.3413 o 34.13%, hacia la derecha o hacia la izquierda, es decir, hacia arriba o hacia abajo de la media.

Distribución normal estándar. Tiene media aritmética de 0 y desviación estándar de 1.

De la misma manera, se sabe que el área contenida bajo la curva y dentro de 2 desviaciones estándar alrededor de la media contiene 95.45% del área total y que el área contenida bajo la curva y dentro de 3 desviaciones estándar alrededor de la media contiene 99.73% del área total. Estos valores del área bajo la curva normal son muy utilizados en estadística y se definen en términos de la desviación estándar. Se les resume en el siguiente cuadro.

Desviaciones estándar alrededor de la media	Área bajo la curva normal
1	64.26
2	95.45
3	99.73

ejercicios 3.2 Medidas de dispersión

Para una serie simple

1. A continuación se muestran la cantidad de tortas que se vendieron en una lonchería durante 5 días. Calcule:

a) El rango.
b) La desviación media.
c) La varianza.
d) La desviación estándar.
e) El coeficiente de variación.

Lunes	Martes	Miércoles	Jueves	Viernes
98	93	95	97	100

2. Una fábrica de golosinas elabora 10 diferentes productos; a continuación se presentan los costos de producción por cada 100 piezas de los diferentes dulces. Calcule:

a) El rango.
b) La desviación media.
c) La varianza.
d) La desviación estándar.
e) El coeficiente de variación.

Bombón	$20
Caramelo	$33
Caramelo suave	$41
Chocolate	$63
Cocada	$38
Mentas	$27
Nuez garapiñada	$63
Oblea	$19
Paleta	$16
Tamarindo	$52

3. A continuación se muestran los salarios mensuales (en miles de pesos) de 15 trabajadores de una planta envasadora. Calcule:

a) El rango.
b) La desviación media.
c) La varianza.
d) La desviación estándar.
e) El coeficiente de variación.

6	6.25	7	8	9.5
6	6.25	7.5	8	10
6.25	6.25	7.5	9	10

4. Se realizó una encuesta en 30 restaurantes de Acapulco, Guerrero, en la que se preguntó el precio de una comida completa por persona. Con los resultados que se muestran en el cuadro siguiente calcule:

a) El rango.
b) La desviación media.
c) La varianza.
d) La desviación estándar.
e) El coeficiente de variación.

100	105	118	143	205	250
101	105	119	150	215	300
102	109	125	150	225	340
102	110	125	185	230	400
105	112	135	189	250	420

5. Se contó el número de autos que pasan durante una hora de la mañana por la primera caseta de la carretera México-Puebla en dirección a Puebla durante 45 días. Calcule:

a) El rango.
b) La desviación media.
c) La varianza.
d) La desviación estándar.
e) El coeficiente de variación.

316	319	376	426	498	561	632	697	724
317	320	377	427	509	573	633	703	737
318	323	395	457	519	578	633	705	742
319	346	398	462	533	605	642	706	751
319	353	417	479	534	621	676	708	763

Para una serie de datos y frecuencias

6. Un taxista registró el número de clientes que tiene por día durante un mes (30 días) y obtuvo los resultados que se muestran en el cuadro siguiente. Calcule:

a) El rango.
b) La desviación media.
c) La varianza.
d) La desviación estándar.
e) El coeficiente de variación.

Núm. de clientes	*f*
8	1
10	5
11	3
13	7
14	6
17	3
18	2
19	2
23	1
Total	30

7. Se realizó un estudio para observar cuántas personas se formaban a cierta hora en los cajeros automáticos del banco MexBanc localizados en 25 establecimientos comerciales elegidos al azar y se obtuvieron los resultados que se muestran en el cuadro siguiente. Calcule:

a) El rango.
b) La desviación media.
c) La varianza.
d) La desviación estándar.
e) El coeficiente de variación.

Personas en fila	*f*
0	3
1	4
2	8
3	6
4	1
5	2
6	1
Total	25

8. En el cuadro siguiente se muestra el número de llamadas que atienden en promedio cada día 55 personas que trabajan en un centro de atención a clientes. Calcule:

a) El rango.
b) La desviación media.
c) La varianza.
d) La desviación estándar.
e) El coeficiente de variación.

Llamadas atendidas	*f*
10	2
12	5
15	8
16	9
18	14
19	13
20	3
21	1
Total	55

9. Los datos siguientes representan el número de veces que una muestra de 160 personas acudió al doctor durante el último año. Calcule:

a) El rango.
b) La desviación media.
c) La varianza.
d) La desviación estándar.
e) El coeficiente de variación.

Núm. de visitas	*f*
1	7
2	9
3	10
5	13
6	16
7	17
9	19
10	23
11	25
14	13
15	8
Total	160

10. Se contó el número de visitantes que asistieron al Museo de Arte Contemporáneo durante 500 días y los resultados se presentan en la siguiente tabla. Calcule:

a) El rango.
b) La desviación media.
c) La varianza.
d) La desviación estándar.
e) El coeficiente de variación.

Núm. de visitantes	*f*
98	4
102	9
115	15
123	19
137	21
148	27

Núm. de visitantes	*f*
150	29
151	33
169	38
179	41
180	43
209	46
227	47
249	44
268	28
294	17
307	11
319	9
321	7
349	5
364	3
372	2
388	1
402	1
Total	500

En una serie de clases y frecuencias

11. Los datos siguientes resumen los resultados de una encuesta en la que se preguntó a 200 niños de 10 años de edad el número de horas que dedican a la semana a ver televisión. Calcule:

a) El rango.
b) La desviación media.
c) La varianza.
d) La desviación estándar.
e) El coeficiente de variación.

Núm. de horas *x*	*f*
4 a menos de 7	13
7 a menos de 10	15
10 a menos de 13	17
13 a menos de 16	26
16 a menos de 19	29
19 a menos de 22	28
22 a menos de 25	21
25 a menos de 28	18
28 a menos de 31	14
31 a menos de 34	10
34 a menos de 37	9
Total	200

12. Se contó en número de usuarios que utilizaron la ruta 2 del transporte interno de Ciudad Universitaria durante 90 días y se obtuvieron los resultados que se muestran en el cuadro siguiente. Calcule:

a) El rango.
b) La desviación media.
c) La varianza.
d) La desviación estándar.
e) El coeficiente de variación.

Núm. de pasajeros	*f*
800 a menos de 1 055	1
1 055 a menos de 1 310	1
1 310 a menos de 1 565	2
1 565 a menos de 1 820	4
1 820 a menos de 2 075	9
2 075 a menos de 2 330	12
2 330 a menos de 2 585	15
2 585 a menos de 2 840	19
2 840 a menos de 3 095	20
3 095 a menos de 3 350	7
Total	90

13. En el cuadro siguiente se muestra el número de clientes formados, a determinada hora, en las 115 sucursales de un banco. Calcule:

a) El rango.
b) La desviación media.
c) La varianza.
d) La desviación estándar.
e) El coeficiente de variación.

Núm. de clientes	*f*
5 a menos de 10	8
10 a menos de 15	11
15 a menos de 20	13
20 a menos de 25	14
25 a menos de 30	15
30 a menos de 35	14
35 a menos de 40	13
40 a menos de 45	11
45 a menos de 50	10
50 a menos de 55	6
Total	115

14. En una distribuidora de motocicletas se vendieron 80 motos de diferentes precios en el último mes. Calcule:

a) El rango.
b) La desviación media.

c) La varianza.
d) La desviación estándar.
e) El coeficiente de variación.

Precio de venta (miles de pesos)	*f*
60 a menos de 67	22
67 a menos de 74	17
74 a menos de 81	15
81 a menos de 88	8
88 a menos de 95	7
95 a menos de 102	5
102 a menos de 109	3
109 a menos de 116	2
116 a menos de 123	1
Total	80

15. Se realizó un estudio para determinar cuánto gastaron en publicidad 65 de las más importantes empresas establecidas en México. Calcule:

a) El rango.
b) La desviación media.
c) La varianza.
d) La desviación estándar.
e) El coeficiente de variación.

Gastos publicitarios (millones de pesos)	*f*
10 a menos de 25	4
25 a menos de 40	7
40 a menos de 55	9
55 a menos de 70	12
70 a menos de 80	15
80 a menos de 95	8
95 a menos de 110	5
110 a menos de 125	3
125 a menos de 140	2
Total	65

Teorema de Chebyshev

16. En la siguiente tabla se muestran los precios de 10 artículos. Calcule la cantidad de datos que se encuentran dentro de $K = 1.5$ desviaciones estándar alrededor de la media, así como el intervalo al que corresponde.

Artículo	Precio
A	$55
B	$120
C	$15
D	$100
E	$230
F	$43
G	$118
H	$258
I	$97
J	$85
Total	1 121

17. Se registra el número de órdenes que se sirven diariamente en un restaurante de comida rápida durante 20 días. Calcule la cantidad de datos que se encuentran dentro de $K = 2$ desviaciones estándar alrededor de la media, así como el intervalo correspondiente.

239	251	265	285	295
240	253	270	287	304
242	255	272	293	310
242	264	278	295	321

18. En la siguiente tabla se muestra cuántas piezas de cada artículo se vendieron durante una semana en una papelería. Calcule la cantidad de datos que se encuentran dentro de $K = 2.5$ desviaciones estándar alrededor de la media, así como el intervalo correspondiente.

Artículo	Piezas vendidas
Bolsa de regalo	5
Cartulina	13
Cinta adhesiva	4
Corrector	7
Cuaderno	4
Goma	11
Lápiz	21
Lápiz adhesivo	9
Monografía	95
Moño	3
Papel lustre	22
Pluma	19
Regla	12
Sacapuntas	10
Tijeras	7
Total	242

19. Se registró el número diario de visitantes a la sección de reptiles de un zoológico durante 2 meses. Calcule la cantidad de datos que se encuentran dentro de $K = 2$ desviaciones estándar alrededor de la media, así como el intervalo correspondiente.

Visitantes	f
30	1
37	1
42	1
52	2
53	2
64	3
67	2
69	4
75	4
88	6
93	9
97	8
112	7
118	6
123	3
138	1
145	1
Total	60

20. En relación con el problema del taxista que registró el número de clientes que tiene por día durante un mes (los datos se reproducen en seguida), calcule la cantidad de datos que se encuentran dentro de $K = 3$ desviaciones estándar alrededor de la media y el intervalo correspondiente.

Núm. de clientes	Días f
8	1
10	5
11	3
13	7
14	6
17	3
18	2
19	2
23	1
Total	30

3.3 Medidas de composición: la proporción

Aunque es una medida muy sencilla y de todos conocida, dada su importancia vale la pena repasarla, aunque sea en forma breve. Prácticamente toda persona, desde temprana edad, tiene idea de lo que es una proporción. Por ejemplo, si se le pregunta a cualquier estudiante de nivel medio para arriba que se ponga de pie en una clase y diga aproximadamente qué proporción de mujeres (o de hombres) hay en su grupo, no le costará ningún esfuerzo hacer una buena aproximación, aun si no sabe la forma exacta de calcularla. La **proporción** se calcula dividiendo el número de casos que tienen la característica de interés entre el total de elementos de la muestra (o de la población) o, en símbolos:

Proporción. Se calcula dividiendo el número de casos que tienen la característica de interés entre el total de elementos de la muestra (o de la población).

$$p = \frac{n_I}{n} \tag{3.28}$$

En donde n_I es el número de casos de interés y n es el total de elementos de la muestra. La proporción para una población, simplemente sustituyendo los símbolos apropiados es:

$$\pi = \frac{N_I}{N} \tag{3.29}$$

Con un ejemplo sencillo basta para ilustrar la idea; si en grupo de 50 estudiantes hay 30 mujeres:

$$p = \frac{n_I}{n} = \frac{30}{50} = 0.6$$

O sea que la proporción de mujeres es del 0.6 (al tanto por uno) o de $0.6 \times 100 = 60\%$, al tanto por ciento.

Un detalle que vale la pena anotar aquí y que se vuelve a tratar en varios capítulos posteriores, empezando con el capítulo 4 "Introducción a la teoría de la probabilidad" es que, cuando se manejan proporciones, básicamente se divide a los elementos de la muestra en 2 grupos. En el ejemplo, son mujeres y hombres y, como ambos grupos conforman la totalidad, o 100% de la muestra, si se sabe que 60% son mujeres, automáticamente se sabe también que los hombres representan 40% restante, para completar 100%. Esto se expresa con símbolos como:

$$p + q = 1, \text{ o despejando la } p\text{: } p = 1 - q$$

Esto mismo expresado en la simbología adecuada para poblaciones:

$$\pi = 1 - q$$

Esta relación se utiliza frecuentemente en estadística, particularmente en lo que se refiere a variables binomiales y a la distribución de probabilidad binomial que se estudia en el capítulo 5 "Distribuciones discretas (discontinuas) de probabilidad".

3.4 Medidas de forma: momentos

En seguida se presentan otras 2 medidas importantes en estadística, las cuales sirven para analizar la forma de una distribución: el apuntamiento, también conocido como *curtosis* y su sesgo. La **curtosis** mide qué tan puntiaguda o qué tan aplanada es una distribución, en tanto que el **sesgo** mide qué tan centrada o simétrica es, o sea, qué tan sesgada es una distribución.

Curtosis. Mide lo aplanado o puntiagudo de una distribución.
Sesgo. Mide lo centrado o simétrico (sesgado) de una distribución.

Sin embargo estas 2 medidas se relacionan con lo que se conoce como *momentos* y éstos, a su vez, con 2 medidas que ya se estudiaron: la media aritmética y la desviación media. Por ello, se comienza por revisar este concepto. El término *momento* proviene del campo de la física porque estas medidas, la media, la desviación media, el sesgo y la curtosis tienen que ver con las nociones de centro de gravedad e inercia en física, aunque no se abunda más aquí sobre este tema.

Se tienen momentos respecto al origen y a la media. El primer momento respecto al origen, 0, se define como:

$$M_{1,0} = \frac{\sum (X_i - 0)^1}{n}$$

Que, simplificando, se puede ver que es la conocida media aritmética:

$$M_{1,0} = \bar{X} = \frac{\sum X_i}{n}$$

Por su parte, el primer momento respecto a la media es

$$M_{1,\bar{X}} = \frac{\sum (X_i - \bar{X})^1}{n} = \frac{\sum (X_i - \bar{X})}{n}$$

Este primer momento respecto a la media es igual a 0 porque, como se vio antes, la sumatoria de esas diferencias entre cada dato y la media siempre es igual a 0:

$$\sum (X_i - \bar{X}) = 0$$

Por lo que el cociente es también igual a 0. Se vio también que la desviación media es el promedio de los valores absolutos de esas diferencias:

$$DM = \frac{\sum |X_i - \bar{X}|}{n}$$

La cual, aunque no es propiamente el primer momento respecto a la media, es una medida que se desprende de ese primer momento.

Estos 2 primeros momentos son precisamente primeros porque dichas diferencias se elevan a la potencia 1 (por ello se anotó el exponente 1 cuando no es necesario hacerlo).

El segundo momento —respecto a la media— se define, entonces, como:

$$M_{2,\bar{X}} = \frac{\sum (X_i - \bar{X})^2}{n} = s^2$$

En donde las diferencias entre las observaciones X_i y la media se elevan a la potencia 2, al cuadrado. Y este segundo momento respecto a la media es, precisamente, la varianza.

Como los demás momentos respecto al origen no son de interés aquí, se limita su consideración a ese primer momento, que es la media aritmética y se simplifica la notación para los momentos respecto a la media, comenzando con el segundo momento, de la siguiente manera:

$$M_2 = \frac{\sum\left(X_i - \bar{X}\right)^2}{n}$$

sin anotar en el subíndice que se trata de un momento respecto a la media.

3.4.1 Tercer momento respecto a la media y el coeficiente de sesgo

Como ya se estudió, el sesgo se refiere a la simetría o, más bien, a la falta de simetría de una distribución, la cual se evalúa en relación con el centro de la distribución, marcado por su media aritmética. En otras palabras, una distribución es simétrica si las 2 partes a la izquierda y a la derecha de la media son reflejo de espejo la una de la otra. Una manera de medir el sesgo es a través del tercer momento respecto a la media:

$$M_3 = \frac{\sum\left(X_i - \bar{X}\right)^3}{n} \tag{3.30}$$

Sin embargo, este tercer momento se da en las unidades originales y, para convertirlo en unidades relativas, se le divide entre el cubo de la desviación estándar, con lo que se obtiene el coeficiente de sesgo:

$$CS = \frac{M_3}{\sigma^3} = \frac{\frac{\Sigma\left(X_i - \bar{X}\right)^3}{n}}{\sigma^3} = \frac{\sum\left(X_i - \bar{X}\right)^3}{n\sigma^3} \tag{3.31}$$

Cuando la distribución es simétrica, este coeficiente de sesgo vale 0, en tanto que, cuando se sesga a la izquierda, el coeficiente es negativo, y tanto más negativo cuanto más sesgada es la distribución. De manera equivalente, cuando la distribución se sesga a la derecha, el coeficiente es positivo y lo es más entre más sesgada a la derecha es la distribución. El sesgo a la izquierda significa que la cola izquierda es más gruesa que la derecha y viceversa.

Como se requieren los cálculos de los cubos de las diferencias entre los datos y su media para el cálculo de este coeficiente de variación y dado que, además, se necesitan las cuartas potencias de estas mismas diferencias para la determinación del coeficiente de curtosis que se analiza en la sección siguiente, se posponen los ejemplos de cutorsis para el final de esta sección para evitar mayores duplicidades en las tablas, ya que se utilizan los mismos datos de los ejemplos anteriores sobre costos de producción, número de personas formadas en una fila y gastos de publicidad.

3.4.2 Cuarto momento respecto a la media y el coeficiente de curtosis o apuntamiento

La comparación de apuntamiento o aplanamiento se hace en comparación con la distribución normal. Así, se dice que una **distribución** es **mesocúrtica** si es tan puntiaguda o aplanada, según sea como la distribución normal, en tanto que una **distribución platicúrtica** tiene un pico relativamente bajo en comparación con la distribución normal lo cual implica, además, que los extremos o colas de la distribución se extienden considerablemente hacia los lados. Por su parte, una **distribución leptocúrtica** es la que tiene un pico prominente al centro (también en comparación con la distribución normal), sus lados bajan en forma marcada y sus extremos no se extienden mucho o, en otras palabras, se extienden menos que en la distribución normal.

Distribución mesocúrtica. Distribución tan puntiaguda o aplanada, según sea, como la distribución normal.
Distribución platicúrtica. Tiene un pico relativamente bajo en comparación con la distribución normal.
Distribución leptocúrtica. Tiene un pico prominente al centro en comparación con la distribución normal.

Una manera de medir esta curtosis es a través del cuarto momento, que se define como:

$$M_4 = \frac{\sum\left(X_i - \bar{X}\right)^4}{n} \tag{3.32}$$

Así como se hizo con el coeficiente de sesgo, se divide este cuarto momento entre la cuarta potencia de la desviación estándar para obtener una medida relativa, que es independiente de las unidades originales, para obtener un coeficiente de curtosis, de la siguiente manera:

$$CK = \frac{M_4}{\sigma^4} - 3 = \frac{\frac{\sum (X_i - \bar{X})^4}{n}}{\sigma^4} - 3 = \frac{\sum (X_i - \bar{X})^4}{n\sigma^4} - 3 \tag{3.33}$$

La razón del "–3" en la fórmula es que el valor de la curtosis para la distribución normal es, precisamente, 3 y, como este apuntamiento se mide respecto a la distribución normal, esta resta del 3 permite interpretar este CK en términos de esta distribución. Así, una curtosis positiva indica una distribución puntiaguda en tanto que una curtosis negativa muestra una distribución aplanada.

En los ejemplos siguientes se ilustran los procedimientos para calcular estas medidas de sesgo y de curtosis para los tres tipos de series o distribuciones que se han venido utilizando: series simples, de datos y frecuencias y de clases y frecuencias.

EJEMPLO 3.31

En una serie simple

En el ejemplo 3.26 se calcularon la varianza y la desviación estándar para los costos de producción por cada 100 piezas de diferentes golosinas. Se reproduce en seguida la tabla con la que se hicieron esos cálculos, y se incluyen las operaciones necesarias para calcular los momentos tercero y cuarto.

Producto	Costo	$X_i - \bar{X}$	$(X_i - \bar{X})^2$	$(X_i - \bar{X})^3$	$(X_i - \bar{X})^4$
Bombón	\$20	–16.5	272.25	–4 492.13	74 120.06
Caramelo	\$33	–3.5	12.25	–42.88	150.06
Caramelo suave	\$41	4.5	20.25	91.13	410.06
Chocolate	\$63	26.5	702.25	18 609.63	493 155.06
Cocada	\$38	1.5	2.25	3.38	5.06
Mentas	\$27	–9.5	90.25	–857.38	8,145.06
Nuez garapiñada	\$56	19.5	380.25	7 414.88	144 590.06
Oblea	\$19	–17.5	306.25	–5 359.38	93 789.06
Paleta	\$16	–20.5	420.25	–8 615.13	176 610.06
Tamarindo	\$52	15.5	240.25	3 723.88	57 720.06
Total	365	0	2 246.5	10 476.00	1 048 694.63

De donde

$$M_3 = \frac{\sum (X_i - \bar{X})^3}{n} = \frac{10\,476}{10} = 1\,047.6$$

Y, ya se calculó antes la desviación estándar,

$$\sigma = \sqrt{\frac{\sum (X_i - \bar{X})^2}{\sum f_i}} = \sqrt{\frac{2\,446.5}{10}} = 15.64$$

De donde

$$CS = \frac{M_3}{\sigma^3} = \frac{\frac{\sum (X_i - \bar{X})^3}{n}}{n\sigma^3} = \frac{10\,476}{10(15.64^3)} = \frac{10\,476}{38\,256.94} = 0.274$$

Este valor indica que esta serie de costos de producción se sesga hacia la derecha.

Se calculan ahora el cuarto momento y el coeficiente de curtosis:

$$M_4 = \frac{\sum (X_i - \bar{X})^4}{10} = \frac{1\,048\,694.63}{10} = 104\,689.46$$

$$CK = \frac{M_4}{\sigma^4} - 3 = \frac{\sum (X_i - \bar{X})^4}{n\sigma^4} - 3$$

$$= \frac{1\,048\,694.63}{10(15.64^4)} - 3 = \frac{1\,048\,694.63}{598\,338.56} - 3 = 1.75 - 3 = -1.25$$

Este valor del coeficiente de curtosis indica que esta serie es más aplanada que una normal.

EXCEL Este paquete tiene 2 funciones que permiten calcular sesgo y curtosis: COEFICIENTE. ASIMETRIA y CURTOSIS.

Para la función COEFICIENTE.ASIMETRIA, la fórmula que usa Excel es:

$$\frac{n}{(n-1)(n-2)} \sum \left(\frac{X_i - \bar{X}}{s} \right)^3$$

Para la CURTOSIS se utiliza:

$$\left\{\frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum\left(\frac{X_i-\bar{X}}{s}\right)^4\right\}-\frac{3(n-1)^2}{(n-2)(n-3)}$$

Como puede observarse, ambas se basan en momentos. Sin embargo, las 2 se calcularon en forma considerablemente distinta a como se propuso aquí.

Utilizando estas funciones con los datos de este ejemplo se obtienen los siguientes resultados:

=COEFICIENTE.ASIMETRIA(A1:A10) = 0.1143
=CURTOSIS(A1:A10) = –1.1839

Como puede verse, el coeficiente de asimetría de 0.1143 que se obtiene es parecido a 0.274 que se obtuvo antes y pareciera conducir a las mismas conclusiones respecto al sesgo de la distribución; y lo mismo sucede con el –1.25 que se obtuvo para el coeficiente de curtosis.

Sin embargo, además de que no se encontró bibliografía que citara las fórmulas de Excel, las fórmulas propuestas aquí son más sencillas y, por ello, no se sugiere la utilización de Excel para estas 2 medidas.[1]

Ejemplo 3.32

En una serie de datos y frecuencias

En el ejemplo 3.27 se realizó un estudio para observar cuántas personas se formaban a cierta hora en los cajeros automáticos del banco MexBanc localizados en 25 establecimientos comerciales y se calculó su varianza, su desviación estándar y su coeficiente de variación. En el cuadro siguiente se reproducen las operaciones que se realizaron ahí, junto con los cálculos necesarios para determinar aquí el coeficiente de sesgo y el de curtosis.

P en fila	*f*	*fX*	$X_i-\bar{X}$	$(X_i-\bar{X})^2$	$f(X_i-\bar{X})^2$	$(X_i-\bar{X})^3$	$f(X_i-\bar{X})^3$	$(X_i-\bar{X})^4$	$f(X_i-\bar{X})^4$
0	3	0	–2.32	5.3824	16.1472	–12.487168	–37.461504	28.9702298	86.9106893
1	4	4	–1.32	1.7424	6.9696	–2.299968	–9.199872	3.03595776	12.143831
2	8	16	–0.32	0.1024	0.8192	–0.032768	–0.262144	0.01048576	0.08388608
3	6	18	0.68	0.4624	2.7744	0.314432	1.886592	0.21381376	1.28288256
4	1	4	1.68	2.8224	2.8224	4.741632	4.741632	7.96594176	7.96594176
5	2	10	2.68	7.1824	14.3648	19.248832	38.497664	51.5868698	103.17374
6	1	6	3.68	13.5424	13.5424	49.836032	49.836032	183.396598	183.396598
Total	25	58			57.44		48.0384		394.957568

Solución:

Se calcula en seguida el coeficiente de sesgo, modificando las fórmulas para una serie de datos y frecuencias:

$$M_3=\frac{\sum f\left(X_i-\bar{X}\right)^3}{n}=\frac{48.0384}{25}=1.92$$

Antes se calculó la desviación estándar

$$\sigma=\sqrt{\frac{\sum f_i\left(X_i-\bar{X}\right)^2}{\sum f_i}}=\sqrt{\frac{57.44}{30}}=1.5$$

De donde

$$CS=\frac{M_3}{\sigma^3}=\frac{\sum f\left(X_i-\bar{X}\right)^3}{n\sigma^3}=\frac{48.0384}{25\left(1.51^3\right)}=\frac{48.0384}{86.074}=0.5581$$

Este valor del coeficiente indica que la distribución se sesga a la derecha.

Y, el cuarto momento y el coeficiente de curtosis,

$$M_4=\frac{\sum f\left(X_i-\bar{X}\right)^4}{n}=\frac{394.96}{25}=15.798$$

[1] Sobre las fórmulas propuestas aquí se puede revisar: Samuel B. Richmond, *Statistical Analysis*, Ronald, Nueva York, 1964.

$$CK = \frac{M_4}{\sigma^4} - 3 = \frac{\sum f\left(X_i - \bar{X}\right)^4}{n\sigma^4} - 3 = \frac{394.96}{25(1.51^4)} - 3 = \frac{394.96}{129.971} - 3 = 3.04 - 3 = 0.04$$

Este coeficiente de curtosis indica que la distribución tiene un apuntamiento prácticamente igual al de una normal.

Ejemplo 3.33

En una serie de clases y frecuencias

En el ejemplo 3.28 se analizó un estudio para determinar cuánto gastaron en publicidad 65 de las más importantes empresas establecidas en México y se calcularon la varianza, la desviación estándar y el coeficiente de variación. En el cuadro siguiente se reproducen las operaciones que se realizaron para encontrar esas medidas y se incluyen los cálculos necesarios para determinar los coeficientes de asimetría y de curtosis.

Gastos	f	Pm	$Pm - \bar{X}$	$(Pm - \bar{X})^3$	$f(Pm - \bar{X})^3$	$(Pm - \bar{X})^4$	$f(Pm - \bar{X})^4$
10 a < 25	4	17.5	–52.62	–145 697.64	–582 790.58	7 666 610.07	30 666 440.26
25 a < 40	7	32.5	–37.62	–53 242.25	–372 695.73	2 002 973.32	14 020 813.25
40 a < 55	9	47.5	–22.62	–11 573.85	–104 164.64	261 800.46	2 356 204.12
55 a < 70	12	62.5	–7.62	–442.45	–5 309.41	3 371.47	40 457.69
70 a < 80	15	77.5	7.38	401.95	6 029.21	2 966.37	44 495.56
80 a < 95	8	92.5	22.38	11 209.35	89 674.76	250 865.15	2 006 921.18
95 a < 110	5	107.5	37.38	52 229.74	261 148.72	1 952 347.80	9 761 739.02
110 a < 125	3	122.5	52.38	143 713.14	431 139.42	7 527 694.34	22 583 083.02
125 a < 140	2	137.5	67.38	305 909.54	611 819.08	20 612 184.76	41 224 369.51
Total	65				334 850.84		122 704 523.62

El tercer momento es:

$$M_3 = \frac{\sum f\left(Pm_i - \bar{X}\right)^3}{n} = \frac{334\ 850.84}{65} = 5\ 151.55$$

Y con la desviación estándar que se calculó antes era:

$$\sigma = \sqrt{\frac{\sum f_i\left(X_i - \bar{X}\right)^2}{\sum f_i}} = \sqrt{\frac{55\ 405.5}{65}} = 29.19$$

Y el coeficiente de sesgo es:

$$CS = \frac{M_3}{\sigma^3} = \frac{\sum f\left(Pm_i - \bar{X}\right)^3}{n\sigma^3} = \frac{334\ 850.84}{65\left(29.19^3\right)}$$

$$= \frac{334\ 850.84}{1\ 616\ 648.64} = 0.207$$

Este valor del coeficiente indica que la distribución se sesga a la derecha.

Y, el cuarto momento y el coeficiente de curtosis,

$$M_4 = \frac{\sum f\left(Pm_i - \bar{X}\right)^4}{n} = \frac{122\ 704\ 523.62}{65} = 1\ 887\ 761.90$$

$$CK = \frac{M_4}{\sigma^4} - 3 = \frac{\sum f\left(Pm_i - \bar{X}\right)^4}{n\sigma^4} - 3 = \frac{122\ 704\ 523.62}{65\left(29.19^4\right)} - 3$$

$$= \frac{122\ 704\ 523.62}{47\ 189\ 973.84} - 3 = 2.06 - 3 = -0.4$$

El cual indica que la distribución es más plana que la normal.

En la tabla 3.22 se resumen estos momentos y coeficientes para los 3 ejemplos (los 3 tipos de series) y en las figuras 3.5, 3.6 y 3.7 se grafican los correspondientes conjuntos de datos, las distribuciones.

Tabla 3.22 Momentos y coeficientes para los 3 ejemplos

Ejemplo / medida	M_3	Coeficiente de sesgo	M_4	Coeficiente de curtosis
Costos de producción	1 047.6	0.274	104 689.46	–1.25
Personas formadas	1.92	0.5581	15.798	0.04
Gastos de publicidad	5 151.55	0.207	1 887 761.90	–0.4

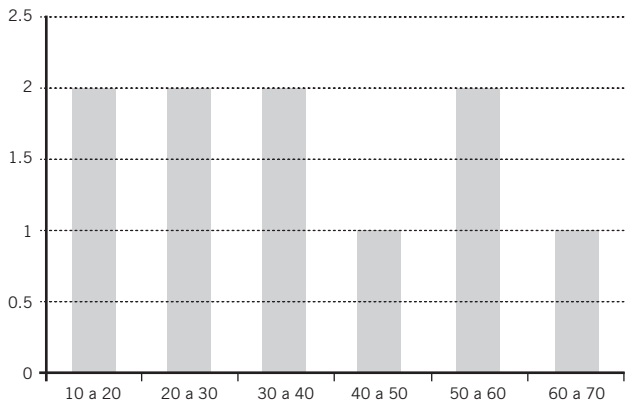


Figura 3.5 La distribución de frecuencias para los datos de costos de producción.

Nótese que, en esta figura 3.5, se agruparon los datos en clases.

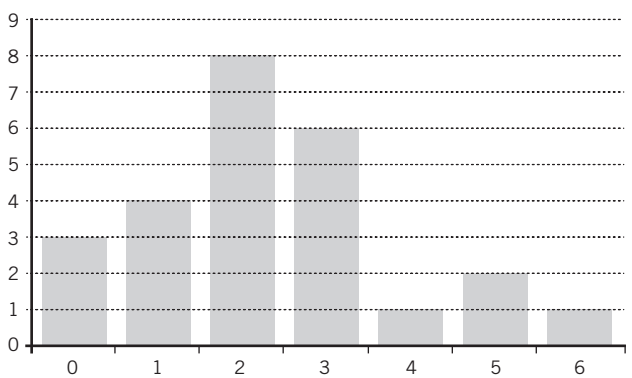


Figura 3.6 La distribución de frecuencias para los datos de personas en una fila.

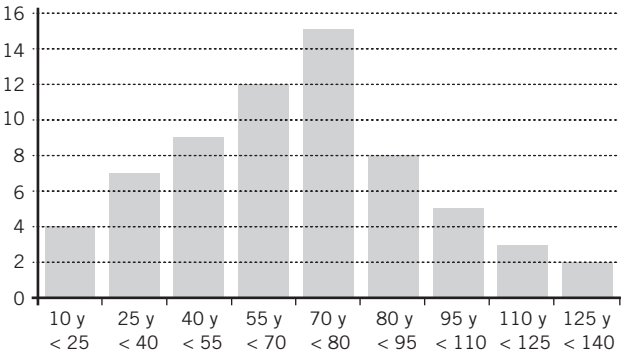


Figura 3.7 La distribución de frecuencias para los datos de gastos de publicidad.

Como la comparación contra una distribución normal resulta difícil, no se puede apreciar claramente el significado de los coeficientes de curtosis. Sin embargo, sí puede verse que la figura 3.5 muestra una distribución aplanada y tiene un coeficiente de curtosis de -1.25, en tanto que pudiera aceptarse que el histograma de la figura 3.6 tiene un apuntamiento equivalente al de la normal ($CS = 0.04$) y que el histograma de la figura 3.6 muestra una distribución ligeramente más aplanada que la normal con un $CS = -0.4$.

Por otra parte, se puede ver en las gráficas cómo los datos parecen extenderse a la derecha al tener las 3 distribuciones coeficiente de sesgo positivo; tiende a verse una cola hacia la derecha; es decir, se sesgan a la derecha, tal y como señala su coeficiente de sesgo. La que parece estar más sesgada corresponde a los datos de las personas que hacen fila que es, precisamente, la serie que tiene el mayor coeficiente de sesgo, 0.5581. Otra manera de visualizar este sesgo es observar que, en estas gráficas sesgadas a la derecha, las observaciones tienden a reunirse más en la parte izquierda de la media. En resumen, el sesgo se interpreta en términos de adónde se extiende la cola de una distribución no simétrica o de hacia dónde se carga la "giba" de la distribución.

ejercicios 3.3 Medidas de forma mediante momentos

Series simples

1. En el ejemplo 3.10 se analizaron las tasas de interés promedio mensual para los Certificados de la Tesorería de la Federación (Cetes) a 28 días, de agosto de 2009 a febrero de 2011. Calcule:

a) El tercer momento y el coeficiente de sesgo.
b) El cuarto momento y el coeficiente de curtosis. Interprételos.

2. En la tabla se presenta la cantidad de minutos que 30 estudiantes invierten para trasladarse de su casa a la escuela. Calcule:

a) El tercer momento y el coeficiente de sesgo.
b) El cuarto momento y el coeficiente de curtosis. Interprételos.

29	20	19	15	43
32	21	31	41	25
25	36	43	23	18
23	15	25	33	19
42	16	15	24	32
28	33	17	32	28

3. En el siguiente cuadro se muestran las temperaturas máxima y mínima (en grados centígrados) para diversas ciudades mexicanas cierto día de febrero. Calcule:

a) El tercer momento y el coeficiente de sesgo.
b) El cuarto momento y el coeficiente de curtosis de las temperaturas mínimas y la de las temperaturas máximas. Interprételos.

Ciudad	Máx.	Mín.
Acapulco	31	23
Aguascalientes	26	14
Campeche	29	22
Cancún	31	24
Chihuahua	32	17
Ciudad Juárez	34	17
Cuernavaca	30	17
Distrito Federal	23	11
Durango	26	11
Guadalajara	27	17
Hermosillo	40	23
La Paz	33	23
Matamoros	34	18
Mazatlán	32	24
Mérida	31	23
Monterrey	31	19
Nuevo Laredo	38	18
Oaxaca	29	18
Puebla	26	14
Puerto Vallarta	32	23
Saltillo	24	10
San Luis Potosí	25	16
Tampico	31	22
Tijuana	29	19
Torreón	31	17
Veracruz	31	22
Villahermosa	31	22

Series de datos y frecuencias

4. Con los datos del ejercicio 2, del tiempo que tardan 30 estudiantes en trasladarse de su casa a su escuela, agrúpelos en una serie de datos y frecuencias. Calcule:

a) El tercer momento y el coeficiente de sesgo.
b) El cuarto momento y el coeficiente de curtosis. Interprételos.

5. Se prueban 400 focos para determinar su vida útil. El resultado se muestra a continuación. Calcule:

a) El tercer momento y el coeficiente de sesgo.
b) El cuarto momento y el coeficiente de curtosis.

Vida útil de los focos (en horas) x	Núm. de focos f
73	11
75	14
77	16
80	19
82	23
85	35
89	41
91	44
94	47
97	39
99	32
100	28
103	21
106	17
107	13
Σ	400

6. En la siguiente tabla se registra el número de clientes que compra determinado número de artículos en un supermercado. Calcule:

a) El tercer momento y el coeficiente de sesgo.
b) El cuarto momento y el coeficiente de curtosis. Interprételos.

Núm. de artículos comprados x	Núm. de clientes f	Núm. de artículos comprados x	Núm. de clientes f
1	19	19	17
2	15	20	30
3	22	21	12
4	19	22	21
5	24	23	18
6	24	24	20
7	17	25	29
8	9	26	12
9	13	27	26
10	26	28	10
11	32	29	16
12	18	30	8
13	11	31	9
14	24	32	11
15	10	33	7
16	27	34	4
17	14	35	3
18	23	Σ	600

Series de clases y frecuencias

7. Convertir los datos de traslados de los 30 estudiantes (ejercicio 2) en una serie de clases y frecuencias. Calcule:

a) El tercer momento y el coeficiente de sesgo.
b) El cuarto momento y el coeficiente de curtosis. Interprételos.

8. En la siguiente tabla se muestran los resultados de una encuesta que se hizo a 100 personas acerca del número de horas por semana que utilizan internet. Calcule:

a) El tercer momento y el coeficiente de sesgo.
b) El cuarto momento y el coeficiente de curtosis. Interprételos.

Núm. de horas que usan internet *x*	Núm. de personas *f*
0 a menos de 5	8
5 a menos de 10	23
10 a menos de 15	38
15 a menos de 20	20
20 a menos de 25	11
Total	100

9. Se preguntó a 150 familias cuánto dinero gastan a la semana en comida rápida, con los resultados que se muestran en la siguiente tabla. Calcule:

a) El tercer momento y el coeficiente de sesgo.
b) El cuarto momento y el coeficiente de curtosis. Interprételos.

x	*f*
80 a 100	9
101 a 121	24
122 a 142	36
143 a 163	48
164 a 184	21
185 a 205	12
Total	150

3.5 Funciones estadísticas de Excel y el complemento "Análisis de datos"

A lo largo del capítulo se mostraron las funciones de Excel que permiten calcular numerosas medidas estadísticas, y en la tabla 3.23 se resumen todas.

Tabla 3.23 Resumen de funciones de Excel que calculan medidas estadísticas

Función	Sintaxis	Descripción de la función
CUARTIL	(matriz,cuartil)	Devuelve el cuartil de un conjunto de datos.
MEDIA.ACOTADA	(matriz,porcentaje)	Devuelve la media de la porción interior de un conjunto de valores de datos.
MEDIA.ARMO	(número1,número2,…)	Devuelve la media armónica de un conjunto de números positivos: el recíproco de la media aritmética de los recíprocos.
MEDIA.GEOM	(número1,número2,…)	Devuelve la media geométrica de una matriz o rango de datos numéricos positivos.
MEDIANA	(número1,número2,…)	Devuelve la mediana o el número central de un conjunto de números.
MODA	(número1,número2,…)	Devuelve el valor más frecuente o que más se repite en una matriz o rango de datos.
PERCENTIL	(matriz,k)	Devuelve al percentil *k*-ésimo de los valores de un rango.
PROMEDIO	(número1,número2,…)	Devuelve el promedio (media aritmética) de los argumentos, los cuales pueden ser números, nombres, matrices o referencias que contengan números.
PROMEDIOA	(ref1,ref2,…)	Devuelve el promedio (media aritmética) de los argumentos; O evalúa el texto como FALSO; 1 como VERDADERO. Los argumentos pueden ser números, nombres, matrices o referencias.

Como se dijo en el capítulo 1, Excel tiene un complemento llamado "Análisis de datos", que comprende un conjunto de rutinas de cálculo para diversos temas esta-

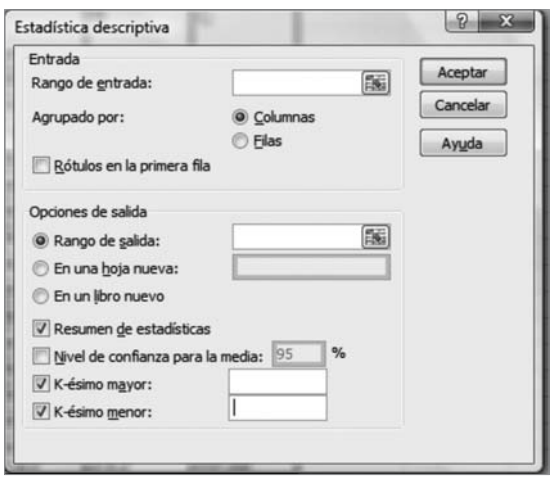
NOTA No se utilizan aquí las funciones Media.Acotada, ni PromedioA.

dísticos. Y este complemento tiene 3 elementos que resultan importantes aquí porque realizan diversos cálculos que se han explicado hasta el momento:

1. *Estadística descriptiva.* Calcula varias de las principales medidas que se tratan en este capítulo.
2. *Media móvil.* Calcula los promedios móviles que se estudiaron en la sección 3.1.5 "Medias o promedios móviles".
3. *Jerarquía y percentil.* Calcula los percentiles que se abordaron en la sección 3.1.8 "Percentiles".

En las subsecciones siguientes se explicará cómo se utilizan estos mecanismos de Excel.

3.5.1 Estadística descriptiva



Para ilustrar el uso de este mecanismo, conviene retomar los datos de los salarios mínimos establecidos en México para el año 2011, de acuerdo con las 3 zonas económicas que se utilizaron en el capítulo anterior para presentar ejemplos de tablas y gráficas. Esos datos se presentaron como ejemplo 2.1 en el capítulo 2 "Presentación de datos: tablas y gráficas" que también se pueden obtener en internet en la dirección: http://www.sat.gob.mx/sitio_internet/asistencia_contribuyente/informacion_frecuente/salarios_minimos/.

Si se colocan esos datos en una hoja de Excel y se selecciona la pestaña "Datos", en el extremo derecho de la cinta de opciones aparece la sección "Análisis" y dentro de ésta se encuentra "Análisis de datos" con su ícono correspondiente. Si se da clic allí, aparecen todos los elementos de análisis, entre los que se encuentra "Estadística descriptiva"; al elegir esta opción, aparece el cuadro de diálogo que se muestra a la izquierda.

Al anotar el rango donde se encuentran los datos en la sección "Rango" y activar la casilla "Resumen de estadísticas", se obtiene el cuadro que se muestra a continuación:

Columna 1		Columna 2		Columna 3	
Media	85.83653	Media	83.61278	Media	81.3319
Error típico	1.955246	Error típico	1.906848	Error típico	1.84893
Mediana	82.69	Mediana	80.7	Mediana	78.46
Moda	81.28	Moda	79.1	Moda	76.91
Desviación estándar	16.59081	Desviación estándar	16.18014	Desviación estándar	15.6887
Varianza de la muestra	275.2551	Varianza de la muestra	261.797	Varianza de la muestra	246.134
Curtosis	27.64019	Curtosis	27.64644	Curtosis	27.6145
Coeficiente de asimetría	5.130427	Coeficiente de asimetría	5.130772	Coeficiente de asimetría	5.12701
Rango	104.55	Rango	101.87	Rango	98.74
Mínimo	74.65	Mínimo	72.8	Mínimo	70.86
Máximo	179.2	Máximo	174.67	Máximo	169.6
Suma	6 180.23	Suma	6 020.12	Suma	5 855.9
Cuenta	72	Cuenta	72	Cuenta	72
Mayor(1)	179.2	Mayor(1)	174.67	Mayor(1)	169.6
Menor(1)	74.65	Menor(1)	72.8	Menor(1)	70.86

Como puede observarse, se tiene una serie de medidas para cada uno de los conjuntos de salarios, correspondientes a las 3 zonas en las que se dividen los salarios mínimos. Algunos detalles que vale la pena resaltar:

Se denomina "Error típico" a lo que en muestreo se conoce como "Error estándar". Se estudia este concepto en el capítulo 7 sobre distribuciones muestrales, mismo que se aplica en varios de los capítulos siguientes que tratan de estadística inferencial: estimación de parámetros y pruebas de hipótesis.

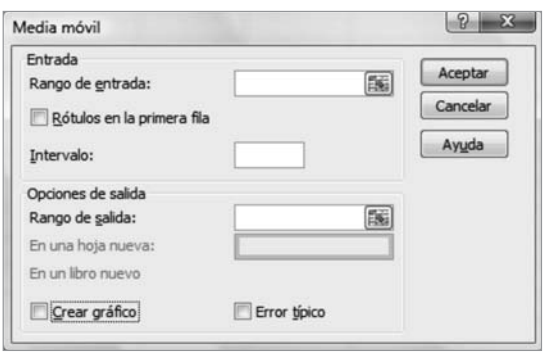
Como Excel considera que los datos provienen de muestras, el cálculo de la varianza y de la desviación estándar se realizan utilizando $n-1$ como denominador.

Los 2 últimos renglones "Mayor(1)" y "Menor(1)" son idénticos a los renglones de arriba identificados como "Máximo" y "Mínimo". Los anteriores "Mayor" y "Menor" son útiles para pedir a Excel que identifique al segundo mayor valor (Mayor(2)) o al tercer menor valor (Menor(3)), por ejemplo.

Activando el cuadro correspondiente a "Nivel de confianza para la media" más un porcentaje, se obtiene una estimación para la media pero, como esto es tema del capítulo 8 que trata, precisamente, de estimación de parámetros, se pospone su uso hasta aquel capítulo.

Finalmente, es importante señalar que, aunque este mecanismo de estadística descriptiva de Excel puede calcular todas esas medidas para varias columnas de datos (como los 3 conjuntos de salarios), si se desean los cálculos para una sola columna es imprescindible colocarlos todos en una sola columna.

3.5.2 Media móvil



En la sección 3.1.5 "Medias o promedios móviles" se explicaron los promedios móviles, como su nombre lo indica, y en el ejemplo 3.9 se calculó un promedio móvil de 5 meses para los datos mensuales de "Flujos netos de Pemex", de agosto de 2008 a febrero de 2011.

Si se colocan estos datos en una hoja de Excel y luego se elige la opción "Media móvil" de "Análisis de datos" que se encuentra en la pestaña "Datos" de la cinta de opciones de Excel, aparece el cuadro de diálogo que se muestra a la derecha.

Si en la sección de "Rango de entrada" se anotan las celdas en donde están los datos, en intervalo se anota 5 y se pone una celda como inicio del rango de salida (por ejemplo C1, si la columna de datos comienza en la celda B1) y, además, se marca el cuadro correspondiente a "Crear gráfico", se obtienen los datos que se muestran en la tabla 3.24.

La tercera columna muestra el promedio móvil calculado por Excel y, como puede verse, las 5 primeras celdas del promedio móvil contienen la anotación "#N/A", lo cual indica error en Excel. No importa en dónde se pida el rango de salida, Excel siempre marca este error que es una deficiencia del propio programa. El resto de los datos, los promedios móviles, son los mismos que se obtuvieron antes en el ejemplo 3.9 inicial.

Tabla 3.24 Media móvil calculada con Excel para los datos de flujo neto de Pemex

Ago. 2008	2 659.50	#N/A
Sep. 2008	3 041.30	#N/A
Oct. 2008	2 541.50	#N/A
Nov. 2008	4.6	#N/A
Dic. 2008	817.7	1 812.92
Ene. 2009	–703.3	1 140.36
Feb. 2009	3 560.20	1 244.14
Mar. 2009	–141.2	707.60
Abr. 2009	690.4	844.76
Mayo 2009	623.2	805.86
Jun. 2009	778.4	1 102.20
Jul. 2009	1 388.20	667.80
Ago. 2009	912.5	878.54
Sep. 2009	2 438.30	1 228.12
Oct. 2009	1 406.70	1 384.82
Nov. 2009	870.6	1 403.26
Dic. 2009	–295.1	1 066.60
Ene. 2010	1 014.70	1 087.04
Feb. 2010	1 870.80	973.54
Mar. 2010	1 852.40	1 062.68
Abr. 2010	475.4	983.64
Mayo 2010	659.4	1 174.54
Jun. 2010	338.5	1 039.30
Jul. 2010	1 900.40	1 045.22
Ago. 2010	1 682.50	1 011.24
Sep. 2010	1 888.70	1 293.90

(*continúa*)

Tabla 3.24 (*continuación*)

Oct. 2010	–973.2	967.38
Nov. 2010	2 202.40	1 340.16
Dic. 2010	3 124.70	1 585.02
Ene. 2011	951.4	1 438.80
Feb. 2011	1 473.00	1 355.66

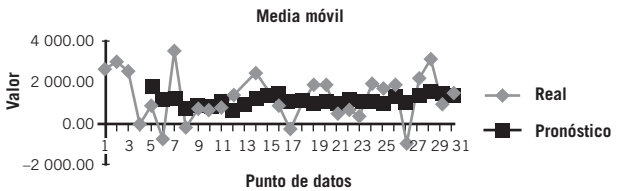
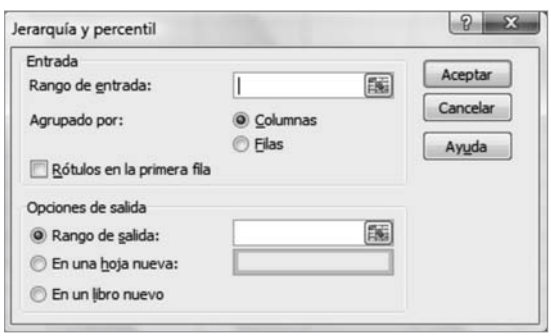


Figura 3.8 Gráfica de media móvil para los datos de flujo neto de Pemex.

Además de los cálculos del promedio móvil, Excel crea la gráfica que se presenta aquí como figura 3.8.

La gráfica que se reproduce es la misma que genera Excel aunque cabe mencionar que es posible mejorarla y ajustarla a los gustos y necesidades del usuario; de lo expuesto hasta aquí sobre este mecanismo de media móvil, es fácil apreciar las ventajas que ofrece para calcular y graficar promedios móviles.

3.5.3 Percentil y jerarquía



En la sección 3.1.8 se revisó el concepto de *percentiles* y la forma en la que se calculan y se ilustraron las operaciones en el ejemplo 3.17, con los datos del número de miles de acciones negociadas de una empresa en 50 días de operaciones. Si se colocan esos datos en una sola columna (en el ejemplo se presentaron en una tabla de 5 renglones y 10 columnas) y se incluye un renglón inicial de encabezado, como "Datos" y se elige "Jerarquía y percentil" de la sección de "Análisis de datos" de la cinta de la pestaña de "Datos" de la cinta de opciones de Excel, se abre el cuadro de diálogo que se muestra del lado izquierdo.

Si se llenan los cuadros necesarios, incluyendo palomear el cuadro "Rótulos en la primera fila" y se anota como rango de salida la celda B1, se obtienen los datos que se muestran ahora en la tabla 3.25.

Tabla 3.25 Resultados de la opción Jerarquía percentiles de Excel con los datos del número de acciones negociadas

Datos	Posición	Columna 1	Jerarquía	Porcentaje
3	50	80	1	100.00
4	49	79	2	97.90
7	48	76	3	95.90
9	46	74	4	91.80
10	47	74	4	91.80
10	45	73	6	89.70
12	44	71	7	87.70
14	43	69	8	85.70
15	41	67	9	81.60
17	42	67	9	81.60
19	40	64	11	79.50
20	39	63	12	77.50
21	38	62	13	75.50
25	37	59	14	73.40
27	35	56	15	69.30
27	36	56	15	69.30
29	34	53	17	67.30

Datos	Posición	Columna 1	Jerarquía	Porcentaje
31	33	52	18	65.30
31	31	48	19	61.20
34	32	48	19	61.20
34	30	47	21	59.10
34	29	45	22	57.10
36	28	43	23	55.10
37	27	39	24	53.00
38	25	38	25	48.90
38	26	38	25	48.90
39	24	37	27	46.90
43	23	36	28	44.80
45	20	34	29	38.70
47	21	34	29	38.70
48	22	34	29	38.70
48	18	31	32	34.60
52	19	31	32	34.60
53	17	29	34	32.60
56	15	27	35	28.50
56	16	27	35	28.50
59	14	25	37	26.50
62	13	21	38	24.40
63	12	20	39	22.40
64	11	19	40	20.40
67	10	17	41	18.30
67	9	15	42	16.30
69	8	14	43	14.20
71	7	12	44	12.20
73	5	10	45	8.10
74	6	10	45	8.10
74	4	9	47	6.10
76	3	7	48	4.00
79	2	4	49	2.00
80	1	3	50	0.00

Al analizar esta tabla se puede observar que ordena los datos de mayor a menor (columna 3) y que señala en la columna 2 la posición que ocupa. Con los datos obtenidos esa información es irrelevante ya que, en el orden original, la posición que ocupa cada dato se obtiene directamente del número de renglón de la hoja de Excel. Por otra parte, la columna de "Jerarquía" es simplemente el orden inverso, lo cual tampoco ofrece nada nuevo. Finalmente, la columna de "Porcentaje" señala la jerarquía porcentual, según señala la opción "Ayuda" del propio programa.

3.6 Resumen

En este capítulo se estudiaron las principales medidas estadísticas; conviene clasificarlas en grupos porque así es más fácil visualizarlas y comprender mejor lo que mide cada una, es decir, lo que dicen a sus usuarios. Las medidas se agrupan en 4 grandes categorías:

1. De posición o medidas de tendencia central.
2. De dispersión.
3. De composición.
4. De forma.

Es posible afirmar que la principal medida de posición es la media aritmética o promedio aritmético, pues, además de que es conocida por prácticamente todos los estudiantes, quizá desde la misma escuela primaria, es de suma utilidad y se utiliza con frecuencia. También se revisaron las medias ponderada, armónica y geométrica y se abundó sobre la relación entre estas 2 últimas y la media aritmética.

Asimismo se estudiaron otras medidas de posición (la moda, la mediana y los percentiles) y su estrecha relación con la mediana, así como la conexión entre la media, la mediana y la moda.

Sobre las medidas de dispersión se revisaron las 2 más importantes en estadística: la varianza y la desviación estándar y se vio *a*) que existe una estrecha relación entre ambas, pues la desviación estándar es simplemente la raíz cuadrada de la varianza y *b*) que ambas miden la dispersión respecto a la media aritmética, ya que los cálculos comienzan por encontrar las diferencias entre cada uno de los datos y la media aritmética, con lo cual se mide qué tan alejados (dispersos) están esos datos de la media aritmética.

Además se examinaron el rango y la desviación intercuartílica, así como 3 aplicaciones comunes de la desviación estándar: el coeficiente de variación, el teorema de Chebyshev y las áreas bajo una curva normal.

En el apartado 3 se revisó la única medida de composición que se incluye: la proporción. Aunque se le dedicó muy poco espacio, se trata en un apartado especial porque es una medida muy importante y su análisis se suele pasar por alto. Aunque la mayoría de las personas tienen una noción considerablemente clara de esta medida desde tempranas edades, resulta importante saber con detalle qué significa y, sobre todo, cómo se calcula.

Finalmente, se revisó el concepto de los momentos respecto al origen y respecto a la media; además se explicó que el segundo momento respecto a la media es, precisamente, la varianza, una medida importante de dispersión que se trató antes. Se analizó también la forma en la que se utilizan los momentos 3 y 4 para construir índices de sesgo y de curtosis, respectivamente, los cuales permiten evaluar la forma de una distribución. El coeficiente de sesgo señala si una distribución es simétrica o si es sesgada, en tanto que el coeficiente de curtosis mide qué tan apuntada o aplanada es una distribución, en comparación con la conocida distribución normal.

3.7 Fórmulas del capítulo

3.1 Técnicas de conteo, permutaciones y combinaciones

3.1.1 Media aritmética

La media en una serie simple:

$$\bar{X} = \frac{\sum X_i}{n} \tag{3.1}$$

La media en una serie de datos y frecuencias:

$$\bar{X} = \frac{\sum f_i X_i}{\sum f_i} = \frac{\sum f_i X_i}{n} \tag{3.2}$$

La media en una serie de clases y frecuencias:

$$\bar{X} = \frac{\sum f_i Pm_i}{\sum f_i} \tag{3.3}$$

3.1.2 Media ponderada

$$\bar{X}_p = \frac{\sum p_i X_i}{\sum p_i} \tag{3.4}$$

3.1.3 Media armónica

$$\bar{X}_a = \frac{1}{\dfrac{\sum_{i=1}^{n} \frac{1}{X_i}}{n}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{X_i}} \tag{3.5}$$

3.1.4 Media geométrica

$$\bar{X}_g = \sqrt[n]{X_1 \cdot X_2 \cdot \ldots \cdot X_n} = (X_1 \cdot X_2 \cdot \ldots X_n)^{\frac{1}{n}} \tag{3.6}$$

3.1.6 Mediana

La mediana en una serie de clases y frecuencias:

$$Med = L_{icmed} + i\left(\frac{\frac{n+1}{2} - f_{aacmed}}{f_{cmed}}\right) \tag{3.7}$$

3.1.7 Moda

$$Mod = L_{ic\,\mathrm{mod}} + i\left(\frac{d_1}{d_1 + d_2}\right) \tag{3.8}$$

3.1.8 Percentiles

Posición que ocupa el percentil:

$$P = (n+1)\frac{p}{100} \tag{3.9}$$

3.1.8.1 Cuartiles

Los cuartiles en una serie de clases y frecuencias:

$$Q_1 = L_{icQ_1} + i\left(\frac{\frac{(n+1)}{4} - f_{aacQ}}{f_{cQ_1}}\right) \tag{3.10}$$

$$Q_2 = Med = L_{icQ_{12}} + i\left(\frac{\frac{n+1}{2} - f_{aacQ_2}}{f_{cQ_2}}\right) \tag{3.11}$$

$$Q_3 = L_{icQ_3} + i\left(\frac{\frac{3(n+1)}{4} - f_{aacQ_3}}{f_{cQ_3}}\right) \tag{3.12}$$

3.2 Medidas de dispersión

3.2.2 Desviación media

La desviación media en una serie simple:

$$DM = \frac{\sum\left|X_i - \bar{X}\right|}{n} \tag{3.13}$$

La desviación media de una serie de datos y frecuencias:

$$DM = \frac{\sum\left|f\left(X - \bar{X}\right)\right|}{\sum f} \tag{3.14}$$

La desviación media en una serie de clases y frecuencias:

$$DM = \frac{\sum\left|f\left(Pm - \bar{X}\right)\right|}{\sum f} \tag{3.15}$$

3.2.3 Desviación intercuartílica

$$DI = Q_3 - Q_1 \tag{3.16}$$

3.2.4 Varianza y desviación estándar

La varianza de una muestra calculada para una serie simple:

$$S^2 = \frac{\sum\left(X_i - \bar{X}\right)^2}{n} \tag{3.17}$$

La varianza de una población calculada para una serie simple:

$$\sigma^2 = \frac{\sum\left(X_i - \mu\right)^2}{n} \tag{3.18}$$

La desviación estándar de una muestra calculada para una serie simple:

$$s = \sqrt{\frac{\sum\left(X_i - \bar{X}\right)^2}{n}} \tag{3.19}$$

La desviación estándar de una población calculada para una serie simple:

$$\sigma = \sqrt{\frac{\sum\left(X_i - \mu\right)^2}{n}} \tag{3.20}$$

La varianza de una muestra calculada para una serie de datos y frecuencias:

$$S^2 = \frac{\sum f\left(X_i - \bar{X}\right)^2}{\sum f_i} \tag{3.21}$$

La varianza de una población calculada para una serie de datos y frecuencias:

$$\sigma^2 = \frac{\sum f\left(X_i - \mu\right)^2}{N} \tag{3.22}$$

La varianza de una muestra calculada para una serie de clases y frecuencias:

$$S^2 = \frac{\sum f_i\left(Pm_i - \bar{X}\right)^2}{\sum f_i} \tag{3.23}$$

La varianza de una población calculada para una serie de clases y frecuencias:

$$\sigma^2 = \frac{\sum f\left(Pm - \mu\right)^2}{N} \tag{3.24}$$

3.2.4.1 Uso de la varianza y la desviación estándar muestrales como estimadores

$$s^2 = \frac{\sum\left(X_i - \bar{X}\right)^2}{n-1} \tag{3.25}$$

$$s = \sqrt{\frac{\sum\left(X_i - \bar{X}\right)^2}{n-1}} \tag{3.26}$$

3.2.5.1 El coeficiente de variación

$$CV = \frac{s}{\bar{X}}(100) \tag{3.27}$$

3.3 Medidas de composición: la proporción

La proporción en una muestra:

$$p = \frac{n_I}{n} \tag{3.28}$$

La proporción en una población:

$$\pi = \frac{N_I}{N} \tag{3.29}$$

3.4 Medidas de forma: momentos

3.4.1 El tercer momento respecto a la media y el coeficiente de sesgo

$$M_3 = \frac{\sum\left(X_i - \bar{X}\right)^3}{n} \tag{3.30}$$

$$CS = \frac{M_3}{\sigma^3} = \frac{\frac{\Sigma\left(X_i - \bar{X}\right)^3}{n}}{\sigma^3} = \frac{\sum\left(X_i - \bar{X}\right)^3}{n\sigma^3} \quad (3.31)$$

3.4.2 El cuarto momento respecto a la media y el coeficiente de curtosis o apuntamiento

$$M_4 = \frac{\sum\left(X_i - \bar{X}\right)^4}{n} \quad (3.32)$$

$$CK = \frac{M_4}{\sigma^4} - 3 = \frac{\frac{\sum\left(X_i - \bar{X}\right)^4}{n}}{\sigma^4} - 3 = \frac{\sum\left(X_i - \bar{X}\right)^4}{n\sigma^4} - 3 \quad (3.33)$$

3.8 Ejercicios adicionales

Media aritmética, mediana, moda y cuartiles para serie simple

1. Se tomó el tiempo de duración de las llamadas (en segundos) de 15 clientes. Calcule:

a) La media aritmética.
b) La mediana.
c) La moda.
d) Los cuartiles.

43	65	68	72	102
51	65	69	93	114
52	65	70	95	136

2. Las horas de estudio que 55 estudiantes dedicaron para la preparación de un examen se muestran a continuación. Calcule:

a) La media aritmética.
b) La mediana.
c) La moda.
d) Los cuartiles.

5	10	14	16	18	20	23	25	27	31	40
5	12	15	16	18	20	24	25	29	35	40
7	12	15	17	18	21	24	26	30	35	42
8	14	15	17	19	22	25	26	31	36	42
9	14	16	18	19	22	25	26	31	36	42

3. El director de recursos humanos de una maquiladora informó que cada uno de los 25 trabajadores de la sección 4 elaboró las siguientes cantidades de pantalones en la última semana. Calcule:

a) La media aritmética.
b) La mediana.
c) La moda.
d) Los cuartiles.

18	21	23	25	28
19	21	24	25	29
19	22	24	26	30
19	22	24	28	33
20	23	24	28	33

4. A continuación se muestran los precios de venta en pesos de 45 refrescos de sabor de una compañía, tomados de una muestra de 45 supermercados de la ciudad de México. Determine:

a) La media.
b) La mediana.
c) La moda.
d) Los cuartiles.

4	4.5	4.5	4.7	5	5	5	5.1	5.3
4	4.5	4.5	4.7	5	5	5	5.1	5.3
4	4.5	4.5	4.8	5	5	5	5.1	5.35
4	4.5	4.5	4.8	5	5	5	5.2	5.4
4	4.5	4.7	4.8	5	5	5	5.2	5.5

Media aritmética, mediana, moda y cuartiles para serie de datos y frecuencias

5. Se presenta a continuación el número de materias reprobadas por 125 estudiantes de la licenciatura en administración, durante su primer semestre. Calcule:

a) La media aritmética.
b) La mediana.
c) La moda.
d) Los cuartiles.

Materias reprobadas	*f*
1	39
2	28
3	19
4	13
5	11
6	8
7	7
Total	125

6. Se preguntó a 80 personas el número de horas a la semana que dedican al aprendizaje de otro idioma. Los resultados se muestran a continuación. Calcule:

a) La media aritmética.
b) La mediana.

c) La moda.
d) Los cuartiles.

Horas a la semana	Frecuencia *f*
2.5	1
3	1
4	1
4.5	1
5	4
7.5	10
8	15
10	18
10.5	13
11	7
12	4
13	2
13.5	1
14	1
15	1
Suma	80

7. Se realizó un estudio para conocer el tiempo que 60 jóvenes de 17 años utilizaban algún tipo de reproductor de música portátil al día; los resultados se muestran en la siguiente tabla. Determine:

a) La media aritmética.
b) La mediana.
c) La moda.
d) Los cuartiles.

Horas por día	Frecuencia *f*
1	1
1.3	1
2	1
2.5	2
3	2
3.3	2
3.5	2
4	4
4.2	4
4.5	7
5	10
5.5	8
6.2	4
6.4	4
7	2
7.5	2
8	1
8.3	1
9	1
10	1
Suma	60

8. Se preguntó a 120 personas el número de años que tardaron en terminar el bachillerato, a continuación se muestran los resultados. Determine:

a) La media aritmética.
b) La mediana.
c) La moda.
d) Los cuartiles.

Años	Frecuencia *f*
2	4
2.5	5
3	73
3.5	19
4	6
4.5	4
5	3
5.5	2
6	1
7	1
7.5	1
8	1
Suma	120

9. Un dentista registró el número de caries que encontró en 45 pacientes suyos; en la siguiente tabla se muestran los resultados. Determine:

a) La media aritmética.
b) La mediana.
c) La moda.
d) Los cuartiles.

Núm. de caries	Frecuencias *f*
0	2
1	4
2	6
4	9
5	8
6	7
7	3
9	2
10	2
11	1
12	1
Suma	45

Media aritmética, mediana, moda y cuartiles para serie de clases y frecuencias

10. En una empresa de mensajería se midió el tiempo (en minutos) requerido para procesar, preparar y enviar 55 paquetes. Calcule:

a) La media aritmética.
b) La mediana.
c) La moda.
d) Los cuartiles.

X	*f*
5 a menos de 9	12
9 a menos de 13	18
13 a menos de 17	14
17 a menos de 21	7
21 a menos de 25	4
Total	55

11. Se registró el número de horas de vuelo de 40 pilotos de una línea aérea; los resultados se muestran en la siguiente tabla. Determine:

a) La media aritmética.
b) La mediana.
c) La moda.
d) Los cuartiles.

X	*f*
5 a menos de 8	1
8 a menos de 11	2
11 a menos de 14	3
14 a menos de 17	5
17 a menos de 20	8
20 a menos de 23	10
23 a menos de 26	7
26 a menos de 29	4
Total	40

12. Se midió el colesterol de 100 miembros de un club deportivo; los resultados se muestran a continuación. Determine:

a) La media aritmética.
b) La mediana.
c) La moda.
d) Los cuartiles.

X	*f*
60 a menos de 90	5
90 a menos de 120	7
120 a menos de 150	15
150 a menos de 180	22
180 a menos de 210	24
210 a menos de 240	12
240 a menos de 270	7
270 a menos de 300	8
Total	100

13. En una planta textil se registró el número de accidentes mensuales que ocurrieron en el transcurso de 5 años; la información se muestra en la siguiente tabla. Determine:

a) La media aritmética.
b) La mediana.
c) La moda.
d) Los cuartiles.

X	*f*
0 a menos de 5	24
5 a menos de 10	10
10 a menos de 15	4
15 a menos de 20	2
Total	40

14. Una empresa realizó un estudio para determinar el costo que significó el "robo hormiga" en sus almacenes durante las quincenas de los últimos 2 años, el resultado se muestra a continuación. Determine:

a) La media aritmética.
b) La mediana.
c) La moda.
d) Los cuartiles.

X (miles de pesos)	*f*
1 a menos de 3	2
3 a menos de 5	10
5 a menos de 7	13
7 a menos de 9	9
9 a menos de 11	7
11 a menos de 13	3
13 a menos de 15	2
15 a menos de 17	2
Total	48

3.1.2 Media ponderada

15. En una tienda de ropa para caballero se vendieron 38 trajes a precio regular de $3 200; para la venta de primavera se bajó el precio a $2 700, con lo que se vendieron 42; en la liquidación de final de temporada se bajó el precio a $1 500 y se vendieron 54. Calcule el precio medio ponderado de los trajes.

16. En un restaurante se sirven 10 platillos diferentes; en la siguiente tabla se muestra el número de veces que se pidió cada uno durante el último mes y su precio correspondiente. Calcule el precio promedio ponderado por platillo.

Platillo	Núm. de ventas *f*	Precio *P*
Albóndigas	23	22
Arroz	30	8
Bistec asado	13	27
Consomé	15	12
Enchiladas	17	25
Espagueti	21	18
Milanesa	9	30
Pechuga asada	11	26
Sopa	29	10
Tortas de papa	16	23
Total	184	

17. En una clase de mercadotecnia se presentan 3 exámenes parciales y 1 al final del curso. Cada examen parcial vale 20% y el final 40%. Si un alumno obtuvo en los exámenes parciales: 8.3, 7.5 y 9.2, y en el final 8.5, ¿cuál sería su promedio final?

18. En la siguiente tabla se muestra el porcentaje de egresados de la carrera de administración de 3 diferentes instituciones educativas. Sin embargo, la población estudiantil no es la misma en todas las escuelas, por lo que los porcentajes no muestran un dato real y es por eso que también se incluye la población estudiantil que tuvo cada generación. Calcule la media ponderada para el porcentaje real de egresados en total.

	Porcentaje *X*	Población *P*
A	33	5 000
B	42	3 500
C	51	1 100

19. Una comercializadora de ropa maneja 5 líneas de producto: damas, caballeros, niños, junior y bebés. En la siguiente tabla se presenta el incremento porcentual en cada una de estas líneas así como la cantidad en miles de pesos de ventas, reportadas al último trimestre del año. Determine el crecimiento porcentual general que tuvo la empresa.

Línea	Porcentaje *X*	Ventas $ *P*
Damas	10.3	1 250.14
Caballeros	4.5	1 534.33
Niños	9.2	2 005.52
Junior	13.4	3 100.33
Bebés	8.5	850.25

3.1.3 Media armónica

20. Hay 8 trabajadores en el área 2 de una fábrica de zapatos, cada uno tarda cierto tiempo en terminar un par de ellos. Calcule el tiempo promedio.

Trabajador	Tiempo (minutos)
A	12
B	15
C	8
D	17
E	21
F	11
G	13
H	15

Media armónica, mediana, moda y cuartiles para serie simple

21. Veinte atletas corren una distancia de 400 metros en cierto número de segundos. Calcule:

a) La media armónica.
b) La mediana.
c) La moda.
d) Los cuartiles.

45	47	53	57	65
45	49	53	58	65
45	49	56	58	67
46	49	56	63	76

3.1.4 Media geométrica

22. En cierto estudio se encontró que en el año 2000 existían 4 000 000 de usuarios de internet en México, y para 2008 ya había 16 000 000. Calcule:

a) El incremento geométrico anual.
b) Rango.
c) Desviación media.
d) Varianza.
e) Desviación estándar.
f) Coeficiente de variación para una serie simple.

23. A continuación se muestran la cantidad de tortas que se vendieron en una lonchería durante 5 días. Calcule:

a) El rango.
b) La desviación media.
c) La varianza.
d) La desviación estándar.
e) El coeficiente de variación.

Lunes	Martes	Miércoles	Jueves	Viernes
98	93	95	97	100

24. A continuación se muestran los salarios (en miles de pesos) de 15 trabajadores de una planta envasadora. Calcule:

a) El rango.
b) La desviación media.
c) La varianza.
d) La desviación estándar.
e) El coeficiente de variación.

6	6.25	7	8	9.5
6	6.25	7.5	8	10
6.25	6.25	7.5	9	10

25. Se realizó una encuesta en 30 estudiantes de 18 años en la que se preguntó cuánto gastan (en pesos) en una salida al cine. Calcule:

a) El rango.
b) La desviación media.
c) La varianza.
d) La desviación estándar.
e) El coeficiente de variación.

100	105	118	143	205	250
101	105	119	150	215	300
102	109	125	150	225	340
102	110	125	185	230	400
105	112	135	189	250	420

26. Se contó el número de autos que pasan por la primera caseta de cobro de una autopista, durante 45 días. Calcule:

a) El rango.
b) La desviación media.
c) La varianza.
d) La desviación estándar.
e) El coeficiente de variación.

316	319	376	426	498	561	632	697	724
317	320	377	427	509	573	633	703	737
318	323	395	457	519	578	633	705	742
319	346	398	462	533	605	642	706	751
319	353	417	479	534	621	676	708	763

27. En la siguiente tabla se muestra el número de acciones (en cientos) de cierta empresa que se negociaron en la BMV diariamente durante un mes. Calcule:

a) El rango.
b) La desviación media.
c) La varianza.
d) La desviación estándar.
e) El coeficiente de variación.

2	9	20	28	33	37
3	11	21	30	35	44
6	13	22	32	35	52
8	14	26	32	36	55
9	16	28	33	37	74

Rango, desviación media, varianza, desviación estándar y coeficiente de variación para una serie de datos y frecuencias

28. Se registra en número de botellas de agua que se venden al día en un supermercado durante 2 meses. Calcule:

a) El rango.
b) La desviación media.
c) La varianza.
d) La desviación estándar.
e) El coeficiente de variación.

Botellas vendidas por día	*f*
20	1
24	2
32	3
39	1
42	4
41	6
43	8
44	9
45	7
52	8
53	10
55	9
57	7
62	5
65	3
66	2
75	1
77	1
82	1
92	1
101	1
Total	90

29. El gerente de una planta ensambladora de computadoras quiere conocer el número de productos defectuosos que se reciben en los embarques del proveedor de tornillos. Se tomó una muestra de 40 cajas de la última entrega; los resultados se muestran a continuación. Calcule:

a) El rango.
b) La desviación media.
c) La varianza.
d) La desviación estándar.
e) El coeficiente de variación.

Tornillos defectuosos por caja	*f*
2	1
3	1
4	2

Tornillos defectuosos por caja	*f*
7	2
8	2
10	3
13	6
15	5
16	4
18	4
20	4
23	3
24	2
25	1
Total	40

30. Se hicieron cortes de caja 56 veces en las cajas registradoras de una boutique de ropa en un centro comercial; los datos obtenidos se concentran en la siguiente tabla. Calcule:

a) El rango.
b) La desviación media.
c) La varianza.
d) La desviación estándar.
e) El coeficiente de variación.

Dinero (miles)	*f*
1.04	1
1.11	1
1.32	2
1.34	3
1.57	4
1.87	4
1.92	6
2.06	6
2.21	5
2.33	4
2.67	4
3.01	3
3.14	3
3.47	2
3.89	2
4.12	2
4.18	1
4.46	1
5.21	1
6.32	1
Total	56

31. Se tomó una muestra de 130 carros que hicieron uso de un estacionamiento controlado y se midió el tiempo en minutos que permanecieron en él. Calcule:

a) El rango.
b) La desviación media.
c) La varianza.
d) La desviación estándar.
e) El coeficiente de variación.

Tiempo	*f*
28	1
33	1
38	1
42	2
51	2
57	3
63	3
64	3
67	4
69	5
71	5
75	6
84	8
86	9
95	10
102	10
112	11
115	11
116	10
123	9
125	7
135	6
142	2
151	1
Total	130

32. El gerente de un centro nocturno registró el número de bebidas que se consumieron en cada una de las 55 mesas durante una noche. Los datos se muestran a continuación. Calcule:

a) El rango.
b) La desviación media.
c) La varianza.
d) La desviación estándar.
e) El coeficiente de variación.

Bebidas por mesa	*f*
4	1
8	1
11	1
12	1
16	1
25	2

(*continúa*)

(*continuación*)

Bebidas por mesa	*f*
26	2
27	3
32	3
33	3
35	4
37	4
41	5
42	7
43	8
47	3
52	2
55	2
62	1
66	1
Total	55

Rango, desviación media, varianza, desviación estándar y coeficiente de variación para una serie de clases y frecuencias

33. En una agencia de autos usados se vendieron en el último mes 80 autos de diferentes precios. Calcule:

a) El rango.
b) La desviación media.
c) La varianza.
d) La desviación estándar.
e) El coeficiente de variación.

Precio de venta (miles de pesos)	*f*
60 a menos de 67	22
67 a menos de 74	17
74 a menos de 81	15
81 a menos de 88	8
88 a menos de 95	7
95 a menos de 102	5
102 a menos de 109	3
109 a menos de 116	2
116 a menos de 123	1
Total	80

34. Se realizó un estudio para medir la velocidad de lectura de una muestra de 100 alumnos de nivel secundaria. Calcule:

a) El rango.
b) La desviación media.
c) La varianza.
d) La desviación estándar.
e) El coeficiente de variación.

Velocidad en minutos	*f*
1 a menos de 3	14
3 a menos de 5	23
5 a menos de 7	31
7 a menos de 9	20
9 a menos de 11	12
Total	100

35. Se tomó una muestra de 200 cajeros automáticos ubicados en el país y se registró la cantidad de efectivo que fue sacado al final del día, los resultados se encuentran a continuación. Calcule:

a) El rango.
b) La desviación media.
c) La varianza.
d) La desviación estándar.
e) El coeficiente de variación.

Retiros en efectivo	*f*
10 000 a menos de 15 000	24
15 000 a menos de 20 000	32
20 000 a menos de 25 000	52
25 000 a menos de 30 000	44
30 000 a menos de 35 000	28
35 000 a menos de 40 000	20
Total	200

36. En un basurero municipal se registró la carga con la que ingresaron 70 camiones en el último día. Calcule:

a) El rango.
b) La desviación media.
c) La varianza.
d) La desviación estándar.
e) El coeficiente de variación.

Toneladas	*f*
2 a menos de 4	6
4 a menos de 6	10
6 a menos de 8	15
8 a menos de 10	14
10 a menos de 12	12
12 a menos de 14	8
14 a menos de 16	5
Total	70

37. Se contó el número de consultas a la base de datos escolar cada hora durante 2 días, los resultados se muestran a continuación. Calcule:

a) El rango.
b) La desviación media.
c) La varianza.
d) La desviación estándar.
e) El coeficiente de variación.

Visitas	f
500 a menos de 1 000	2
1 000 a menos de 1 500	6
1 500 a menos de 2 000	8
2 000 a menos de 2 500	12
2 500 a menos de 3 000	11
3 000 a menos de 3 500	7
3 500 a menos de 4 000	1
4 000 a menos de 4 500	1
Total	48

Teorema de Chebyshev en series simples

38. Se registró el número de los ejemplares que se venden de un periódico en una semana. Calcule qué cantidad de datos se encuentran en $K = 2.5$ desviaciones estándar alrededor de la media, así como el intervalo al que corresponde.

Día	Núm. de ejemplares vendidos
Lunes	520
Martes	630
Miércoles	550
Jueves	612
Viernes	618
Sábado	302
Domingo	287

Teorema de Chebyshev en series de datos y frecuencias

39. La cafetería de una escuela registró el número de tazas de café que vendió diariamente durante 30 días. Calcule qué cantidad de datos se encuentran dentro de $K = 1.8$ desviaciones estándar alrededor de la media, así como el intervalo al que corresponde.

Tazas vendidas	f
47	1
49	1
50	1
56	2
63	2
64	3
65	3
68	4
69	5
74	2
77	2
81	2
85	1
93	1
Total	30

41. En relación con el problema en el que se preguntó a 55 personas que trabajan en un centro de atención telefónica el número de llamadas atendidas por día, cuyos datos se reproducen en seguida, calcule la cantidad de datos que se encuentran dentro de $K = 1.2$ desviaciones estándar alrededor de la media, así como el intervalo correspondiente.

Llamadas atendidas	Días f
10	2
12	5
15	8
16	9
18	14
19	13
20	3
21	1
Total	55

Teorema de Chebyshev en series de clases y frecuencias

42. En la siguiente tabla se muestran los gastos de mantenimiento de 60 de los camiones repartidores de una empresa comercializadora de colchones. Calcule la cantidad de datos que se encuentran dentro de $K = 2$ desviaciones estándar alrededor de la media, así como el intervalo al que corresponde.

Gastos	f
5 000 a menos de 7 000	6
7 000 a menos de 9 000	11
9 000 a menos de 11 000	16
11 000 a menos de 13 000	13
13 000 a menos de 15 000	9
15 000 a menos de 17 000	5
Total	60

43. Se contó el número de veces que fue utilizado un cajero automático ubicado en un centro comercial durante 90 días. Calcule la cantidad de datos que se encuentran dentro de $K = 1.8$ desviaciones estándar alrededor de la media, así como el intervalo correspondiente.

Núm. de veces que se utilizó	f
5 a menos de 15	3
15 a menos de 25	5
25 a menos de 35	7
35 a menos de 45	9
45 a menos de 55	16

(*continúa*)

(*continuación*)

Núm. de veces que se utilizó	*f*
55 a menos de 65	18
65 a menos de 75	11
75 a menos de 85	10
85 a menos de 95	6
95 a menos de 105	5
Total	90

44. Se registró el número de usuarios que ingresaron a una sala de *chat* en internet, por hora, durante todo un día. Calcule qué cantidad de datos se encuentran dentro de $K = 1.6$ desviaciones estándar alrededor de la media y el intervalo correspondiente.

Núm. de usuarios	Horas *f*
0 a menos de 25	4
25 a menos de 50	1
50 a menos de 75	2
75 a menos de 90	4
90 a menos de 115	5
115 a menos de 130	3
130 a menos de 155	3
155 a menos de 180	1
180 a menos de 205	1
Total	24

45. En relación con el problema en que se preguntó a 200 niños el número de horas que veían televisión a la semana, cuyos datos se reproducen en seguida, calcule qué cantidad de datos se encuentran dentro de $K = 2.1$ desviaciones estándar alrededor de la media y el intervalo correspondiente.

Núm. de horas *X*	Niños *f*
4 a menos de 7	13
7 a menos de 10	15
10 a menos de 13	17
13 a menos de 16	26
16 a menos de 19	29
19 a menos de 22	28
22 a menos de 25	21
25 a menos de 28	18
28 a menos de 31	14
31 a menos de 34	10
34 a menos de 37	9
Total	200

46. En relación con el problema en el que se registró el número de personas haciendo fila en las sucursales de un banco, calcule qué cantidad de datos se encuentran dentro de $K = 2$ desviaciones estándar alrededor de la media, así como el intervalo correspondiente. Los datos se reproducen en seguida.

Núm. de clientes	*f*
5 a menos de 10	8
10 a menos de 15	11
15 a menos de 20	13
20 a menos de 25	14
25 a menos de 30	15
30 a menos de 35	14
35 a menos de 40	13
40 a menos de 45	11
45 a menos de 50	10
50 a menos de 55	6
Total	115

Momentos y medidas de forma en series simples

47. Se aplica un examen de aptitud a 50 aspirantes al puesto de gerente, se muestran a continuación las calificaciones finales. Calcule:

a) El tercer momento y el coeficiente de sesgo.
b) El cuarto momento y el coeficiente de curtosis e interprételos.

25	43	47	59	67	74	78	85	89	94
32	43	53	63	67	75	79	86	89	94
36	43	54	65	67	75	79	86	90	97
39	45	56	65	68	77	85	87	91	97
42	46	57	66	69	77	85	87	93	97

47. En la siguiente tabla se muestran las propinas que recibió un mesero durante el día. Calcule:

a) El tercer momento y el coeficiente de sesgo.
b) El cuarto momento y el coeficiente de curtosis e interprételos.

5	10	11	13	15
8	10	12	14	15
10	10	12	14	17

48. En una tienda en línea se seleccionaron 25 horas al azar y se registró el número de usuarios conectados. Calcule:

a) El tercer momento y el coeficiente de sesgo.
b) El cuarto momento y el coeficiente de curtosis e interprételos.

2	6	25	65	108
2	7	32	73	114
2	10	41	81	128
5	15	50	89	137
5	17	58	96	149

49. A continuación se muestran las unidades de un nuevo modelo de automóvil que se ensamblaron en una planta durante los últimos 10 días. Calcule:

a) El tercer momento y el coeficiente de sesgo.
b) El cuarto momento y el coeficiente de curtosis.

8	9	9	9	10
8	9	9	10	11

50. En la siguiente tabla se muestra el número de entregas que hace cada uno de los 12 repartidores de una pizzería. Calcule:

a) El tercer momento y el coeficiente de sesgo.
b) El cuarto momento y el coeficiente de curtosis e interprételos.

9	10	10	12
9	10	11	12
9	10	11	13

Momentos y medidas de forma en series de datos y frecuencias

51. Se preguntó a 300 personas el número de libros que leyó durante el último año, los resultados se muestran en la siguiente tabla. Calcule:

a) El tercer momento y el coeficiente de sesgo.
b) El cuarto momento y el coeficiente de curtosis e interprételos.

Núm. de libros leídos *X*	Núm. de personas *f*
0	30
1	39
2	48
3	46
4	42
5	31
6	25
7	11
8	7
9	9
10	6
11	4
12	2
Σ	300

52. Se preguntó a 45 estudiantes el número de días a la semana que practican algún deporte; los resultados se muestran en la siguiente tabla. Calcule:

a) El tercer momento y el coeficiente de sesgo.
b) El cuarto momento y el coeficiente de curtosis e interprételos.

Días *X*	Estudiantes *f*
0	6
1	11
2	6
3	13
4	9
Σ	45

53. Se realizó una encuesta en la cual se preguntó a 330 el número de veces que utilizó el servicio de transporte público durante la última semana. Con los resultados que se muestran en la siguiente tabla calcule:

a) El tercer momento y el coeficiente de sesgo.
b) El cuarto momento y el coeficiente de curtosis e interprételos.

X	*f*
0	4
3	16
4	20
5	30
8	36
10	35
11	33
14	35
16	31
17	29
22	20
24	14
25	10
29	5
33	3
37	2
38	2
39	2
43	1
47	1
50	1
Σ	330

54. Se registró el monto de 100 multas pagadas en una biblioteca pública por atrasos en la devolución de materiales en préstamo a domicilio; en el cuadro siguiente se muestran los datos que se obtuvieron. Calcule:

a) El tercer momento y el coeficiente de sesgo.
b) El cuarto momento y el coeficiente de curtosis e interprételos.

X	f
5	14
6	12
7	10
9	8
10	8
12	7
15	7
25	6
27	6
33	4
34	4
35	3
47	3
52	3
64	2
65	2
66	1
Σ	100

55. En un edificio hay máquinas expendedoras de dulces; se registró el número de productos que se vendieron diariamente durante un mes. Calcule:

a) El tercer momento y el coeficiente de sesgo.
b) El cuarto momento y el coeficiente de curtosis e interprételos.

X	f
8	1
9	1
12	2
13	3
14	3
17	5
21	4
22	2
24	2
27	2
31	1
32	1
33	1
37	1
39	1
Σ	30

Momentos y medidas de forma en series de clases y frecuencias

56. Se contó, durante 45 días, el número de visitas a una página web en determinada hora; los resultados se muestran en seguida. Calcule:

a) El tercer momento y el coeficiente de sesgo.
b) El cuarto momento y el coeficiente de curtosis e interprételos.

X	f
80 a menos de 90	3
90 a menos de 100	6
100 a menos de 110	5
110 a menos de 120	10
120 a menos de 130	8
130 a menos de 140	7
140 a menos de 150	4
150 a menos de 160	2
Total	45

57. En una fábrica ensambladora de juguetes se tomó el tiempo, en minutos, que tardan 60 trabajadores en armar un carrito; en la tabla que aparece en seguida se muestran los resultados. Calcule:

a) El tercer momento y el coeficiente de sesgo.
b) El cuarto momento y el coeficiente de curtosis e interprételos.

X	f
8 a 10	8
11 a 13	16
14 a 16	16
17 a 19	14
20 a 22	6
Total	60

58. En la siguiente tabla se muestran los gastos en servicio teléfono en que incurrieron 65 hogares en el último bimestre. Calcule:

a) El tercer momento y el coeficiente de sesgo.
b) El cuarto momento y el coeficiente de curtosis.

X (cientos de pesos)	f
1.5 a menos de 3	9
3 a menos de 4.5	18
4.5 a menos de 6	16
6 a menos de 7.5	14
7.5 a menos de 9	1
9 a menos de 10.5	7
Total	65

59. El gerente de producción realizó un estudio para determinar los ajustes pertinentes en la línea de montaje, para ello se registró el número de productos defectuosos que se devolvieron a la planta en cada uno de los 75 embarques que salieron este último mes. Calcule:

a) El tercer momento y el coeficiente de sesgo.
b) El cuarto momento y el coeficiente de curtosis e interprételos.

X	*f*
1 a menos de 6	19
6 a menos de 11	17
11 a menos de 16	15
16 a menos de 21	13
21 a menos de 26	11
Total	75

60. A continuación se muestran los gastos en gasolina que mensualmente tuvo una empresa durante los últimos 2 años. Calcule:

a) El tercer momento y el coeficiente de sesgo.
b) El cuarto momento y el coeficiente de curtosis e interprételos.

X (miles de pesos)	*f*
5 a menos de 15	5
15 a menos de 25	9
25 a menos de 35	7
35 a menos de 45	3
Total	24

capítulo 4

Introducción a la teoría de la probabilidad

Sumario

4.1 Teoría de conjuntos y teoría de la probabilidad
4.2 Conceptos básicos, terminología y notación
4.2.1 Conceptos importantes
4.3 Técnicas de conteo, permutaciones y combinaciones
4.4 Interpretaciones de la probabilidad
4.4.1 Interpretación teórica o clásica
4.4.2 La probabilidad como frecuencia relativa
4.4.3 Interpretación subjetiva de la probabilidad
4.5 Axiomas de la probabilidad
4.5.1 Axioma sobre los posibles valores de la probabilidad
4.5.2 Axioma sobre la suma de las probabilidades de los eventos de un espacio muestral
4.5.3 Axioma sobre la probabilidad de ocurrencia de dos o más eventos mutuamente excluyentes
4.6 Regla de la suma de probabilidades
4.7 Probabilidad condicional
4.8 Independencia estadística
4.9 Regla de la multiplicación de probabilidades
4.9.1 La regla de la multiplicación para eventos independientes
4.10 Regla de Bayes
4.11 Resumen
4.12 Fórmulas del capítulo
4.13 Ejercicios adicionales

La probabilidad es un concepto que la mayor parte de las personas comprende intuitivamente. Por ejemplo, casi todas las personas saben que la probabilidad de ganar o perder una apuesta con el lanzamiento de una moneda es de 50%. En otras palabras, al lanzar una moneda, existe la misma probabilidad de que caiga hacia arriba cualquiera de sus 2 lados.

En estadística se analizan situaciones inciertas, como las posibles características de una población con base en una muestra, por lo cual es necesario estudiar la probabilidad en forma sistemática y se puede comenzar por tratar de definir la **probabilidad** como una medida cuantitativa de la posibilidad de ocurrencia de un evento incierto.

Probabilidad. Medida cuantitativa de la posibilidad de ocurrencia de un evento incierto.

En la definición anterior se distinguen los siguientes elementos:

- Es una *medida cuantitativa*. En el lenguaje cotidiano se dice que "la probabilidad de ganar una apuesta con el lanzamiento de una moneda es de 50%". Dicho porcentaje es la medida cuantitativa expresada en tanto por ciento. Como se verá más adelante, en los cálculos se utiliza la expresión del tanto por uno; es decir, se diría que la probabilidad es de 0.50.
- La *posibilidad de ocurrencia* se refiere a que el evento de interés puede ocurrir o no ocurrir ya que la cara de la moneda que se eligió puede caer hacia arriba o puede caer hacia abajo. Se refiere igualmente a que esa posibilidad (probabilidad) puede ser grande o pequeña.
- Al hablar de un **evento** se hace referencia a un suceso o hecho de interés. En el ejemplo, el evento de interés es la cara de la moneda que queda hacia arriba. En algún otro caso, el evento de interés podría ser que un artículo esté o no defectuoso, o que el precio de una acción que se negocia en la bolsa de valores suba, baje o permanezca igual.

Evento. Suceso o hecho de interés.

- *Incierto*. La incertidumbre es, por supuesto, un aspecto inherente a la probabilidad. Esta incertidumbre se debe, en muchos casos, a que se trata de una situación aleatoria (a la suerte, al azar).

Así, la **teoría de la probabilidad**, que es parte de la estadística, se ocupa de analizar la forma en la que se miden diversos sucesos aleatorios. En las 3 secciones siguientes se revisan diversos conceptos que es necesario manejar al analizar probabilidades y en las secciones restantes se analizan los temas básicos de la teoría de la probabilidad, como las interpretaciones que de ella se hacen, sus axiomas, reglas y diversas aplicaciones.

Teoría de la probabilidad. Se ocupa de analizar la forma en la que se miden diversos sucesos aleatorios.

4.1 Teoría de conjuntos y teoría de la probabilidad

Es común la utilización de algunos elementos de la teoría de conjuntos para analizar probabilidades y representar conjuntos de datos, para lo cual son especialmente útiles los símbolos de unión ($\cup$) e intersección ($\cap$) de conjuntos, así como los diagramas de Venn-Euler o diagramas de Venn.

Ejemplo 4.1

En un conjunto de 100 trabajadores hay 50 con estudios universitarios, 35 que son casados y 20 son casados con estudios universitarios. Se puede representar este conjunto de trabajadores mediante el diagrama de Venn de la figura 4.1:

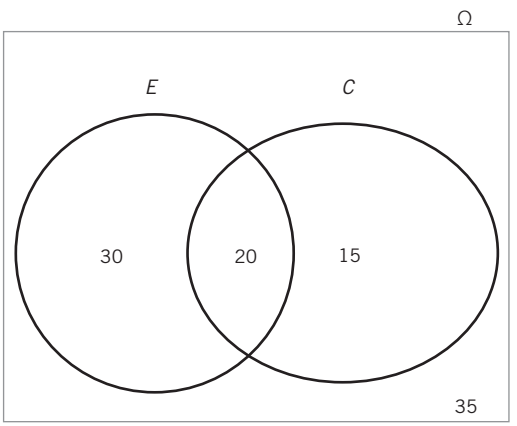


Figura 4.1 Datos del ejemplo 4.1.

En este diagrama se puede ver que 20 trabajadores están casados y tienen estudios universitarios, 30 tienen estudios universitarios pero no están casados, 15 son casados y no tienen estudios universitarios, mientras que 35 no son casados ni tienen estudios universitarios.

Con esta misma información se podría plantear la probabilidad de elegir al azar a uno de estos trabajadores y que fuera o casado (C) o con estudios universitarios (E), lo cual se puede plantear simbólicamente como $P(C \cup E)$, en contraposición con la probabilidad de que el elegido sea casado y que tenga estudios universitarios, lo cual se plantea como $P(C \cap E)$.

Se utiliza una letra n minúscula para representar el número de elementos de un conjunto. Así, el número de elementos del conjunto "Casados" se puede representar como:

$$n(C) = 35$$

Como puede verse en la figura 4.1, la intersección de los conjuntos E y C, que representa al conjunto de personas que tienen estudios universitarios y que son casados, también son 20. Representando esta información con simbología de teoría de conjuntos se tiene lo siguiente:

$$n(E \cap C) = 20$$

Por su parte, el conjunto de personas que tienen estudios universitarios o que son casados, es decir, que tienen cuando menos una de las 2 características son 65, y se representan por la unión de los 2 conjuntos. En símbolos:

$$n(E \cup C) = 65$$

Así, se puede apreciar que el símbolo de la unión se utiliza para denotar lo que en el lenguaje cotidiano expresamos mediante la conjunción "o"; de la misma manera se utiliza el símbolo de la intersección para representar la conjunción "y".

Otro concepto importante de la teoría de conjuntos que conviene recordar es el de **complemento**. Se dice que el complemento de un conjunto A es el conjunto que está formado por todos los elementos que no pertenecen a ese conjunto. De manera que el complemento del conjunto de las personas que tienen educación universitaria está constituido por todas las personas del universo que no tienen estudios universitarios y, como se puede apreciar en la figura 4.1, son 50. En símbolos:

> **Complemento.** El complemento de un conjunto A es el conjunto que se forma por todos los elementos que no pertenecen a ese conjunto.

$$n(E') = 50$$

Una manera común de denotar el complemento de un conjunto es utilizando un apóstrofo después de la letra que lo representa, tal como se hizo en la expresión anterior.

Aquí vale la pena recordar y hacer notar que no es lo mismo el conjunto E que su número de elementos: el conjunto E son todas las personas que tienen estudios universitarios en tanto que su número es 50.

Abundando sobre el concepto de complemento, se puede ver en la figura 4.1 que la intersección de los conjuntos E y C, que es el conjunto de todas las personas que tienen al mismo tiempo educación universitaria y son casados, tiene 20 elementos y que su complemento, que es el conjunto de personas que no tienen ambas características, tiene 80. En símbolos:

$$n(E \cap C)' = 80$$

El complemento de la unión también tiene 30 elementos:

$$n(E \cup C)' = 35$$

Otro concepto importante de la teoría de conjuntos es el de **subconjunto**. Formalmente, se dice que un conjunto B es subconjunto de A, si todos los elementos de B son también elementos de A. Entonces, para nuestro conjunto de 100 personas el conjunto $E \cap C$ es subconjunto tanto del conjunto E como del conjunto C. En símbolos esto se expresa como:

> **Subconjunto.** Un conjunto B es subconjunto de A, si todos los elementos de B son también elementos de A.

$$(E \cup C) \subset E, \text{ y}$$
$$(E \cup C) \subset C$$

En donde el símbolo $\subset$ se lee como "está contenido en".

También, es fácil notar que todo conjunto es subconjunto de sí mismo, $A \subset A$, y que todos los conjuntos son subconjun-

tos de su correspondiente universo, $A_i \subset \Omega$. En donde la letra griega omega (Ω) representa el universo. Otro ejemplo: el conjunto de las personas que no tienen estudios universitarios o que no son casadas, cuyo número es 35, es un subconjunto del universo de 100 personas del ejemplo. En símbolos:

$$(E \cup C)' \subset U$$

EJEMPLO 4.2

Recordando las tablas de contingencias o de doble entrada que se vieron en el capítulo 2, se presenta en seguida la tabla 4.1, que muestra la situación conyugal y el sexo de la población mexicana de 12 años de edad o más, según datos del censo poblacional del año 2010.

Tabla 4.1. Estado conyugal y sexo de la población mexicana de 12 años de edad o más

Situación conyugal	Hombres	Mujeres	Total
Solteros	15 460 577	14 392 540	29 853 117
Casados	17 067 461	17 353 462	34 420 923
Unidos	6 045 370	6 185 310	12 230 680
Separados	970 996	2 211 430	3 182 426
Divorciados	433 354	813 202	1 246 556
Viudos	819 019	2 914 338	3 733 357
No especificado	151 095	109 314	260 409
Total	40 947 872	43 979 596	84 927 468

Fuente: INEGI, disponible en: http://www3.inegi.org.mx/sistemas/TabuladosBasicos/Default.aspx?c=27302&s=est, consultado el 29 de marzo de 2011.

Con los datos de esta tabla se pueden proponer ejemplos adicionales de uniones, intersecciones y complementos de conjuntos, utilizando las iniciales de cada categoría (o 2 letras) para identificar a los diferentes conjuntos implícitos en esos datos: H(ombres), M(ujeres), So(lteros), C(asados), U(nidos), Se(parados), D(ivorciados) y V(iudos) y se puede ver fácilmente en los totales de renglón y de columna el número de elementos de cada uno de estos conjuntos; por otra parte, se pueden identificar subconjuntos formados por uniones e intersecciones de estos conjuntos. Por ejemplo:

$$n(H) = 40\,947\,872$$
$$n(C) = 34\,420\,923$$
$$n(\Omega) = 84\,927\,468$$
$$n\,(U \cap M) = 6\,185\,310$$
$$n\,(U \cap M) \cup (C \cap M) = 6\,185\,310 + 17\,353\,462 = 23\,538\,772$$

EJEMPLO 4.3

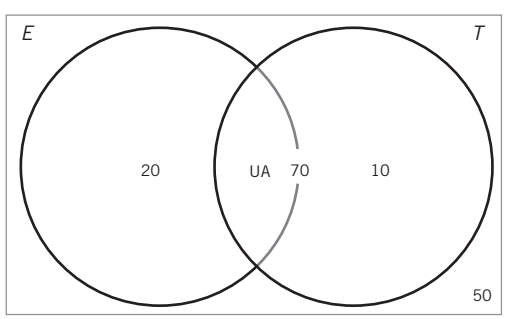
De 150 candidatos que presentaron una solicitud para ocupar el puesto de consultor en informática de una empresa, 90 contaban con experiencia laboral, 80 tenían título profesional y 70 tenían tanto experiencia como título.

A continuación:

a) Represente los subconjuntos mediante un diagrama de Venn.
b) Indique cuántos candidatos:
  i) tienen experiencia,
  ii) tienen título profesional,
  iii) tienen experiencia y título profesional y
  iv) tienen experiencia o título profesional.

Solución:

a)



b)

i) $n(E) = 20$
ii) $n(T) = 10$
iii) $n(S \cap F) = 70$
iv) $n(S \cup F) = 30$

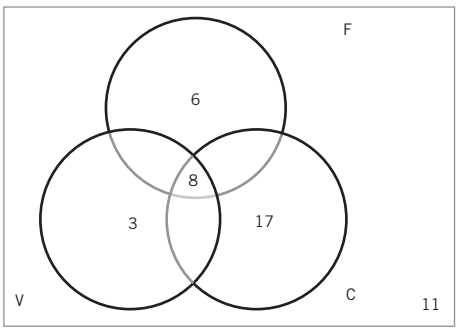
EJEMPLO 4.4

Se realizó una encuesta entre 45 niños de 8 años de edad, en la que se preguntó cuál de los 3 sabores clásicos de helado (fresa, vainilla, chocolate) preferían. Se encontró que 25 preferían chocolate, 14 se inclinaban por fresa, 11 elegían vainilla y 8 gustaban de los 3 sabores.

a) Represente los subconjuntos mediante un diagrama de Venn.
b) Indique cuántos niños prefirieron:
  i) los 3 sabores,
  ii) chocolate o fresa,
  iii) fresa o vainilla o chocolate.

Solución:

a)



b)
  i) $n(F \cap V \cap C) = 8$
  ii) $n(C \cup F) - n(C \cap F \cap V) = (6 + 17 + 8) - 8 = 23$
  iii) $n(F \cup V \cup C) = 26$

EJEMPLO 4.5

En la siguiente tabla se presentan los resultados de 500 entrevistas realizadas a compradores de valores, clasificados por sexo y tipo de valores adquiridos. Indique cuántos compradores:

a) Son hombres.
b) Adquirieron valores de inversión.
c) Adquirieron valores de especulación y son mujeres.
d) Adquirieron valores de inversión y son mujeres, o adquirieron valores de especulación y son hombres.

Tipo de valores	Hombres	Mujeres	Total
De especulación	150	250	400
De inversión	50	50	100
Total	200	300	500

Solución:

a) $n\,(H) = 200$
b) $n\,(\textit{inversión}) = 100$
c) $n\,(\textit{especulación} \cap M) = 250$
d) $n\,((\textit{inversión} \cap M) \cup (\textit{especulación} \cap H) = 50 + 150 = 200$

EJERCICIOS 4.1 Teoría de conjuntos y teoría de la probabilidad

1. Se preguntó a 150 mexicanos qué les gustaba más, viajar a algún destino nacional o viajar al extranjero; 30 dijeron que destino nacional; 73, al extranjero; y 24 ambas opciones.

a) Represente los subconjuntos mediante un diagrama de Venn.
b) ¿Cuántas personas prefieren viajar exclusivamente al extranjero?
c) ¿Cuántas personas prefieren exclusivamente un destino nacional?

2. En una encuesta se preguntó a 100 jóvenes cuáles deportes practicaban y se obtuvieron los siguientes resultados: 35 personas practican béisbol, 15 basquetbol, 40 futbol; 3 sólo béisbol y basquetbol, 10 sólo béisbol y futbol, 2 sólo basquetbol y futbol, y 5 los 3 deportes.

a) Represente los subconjuntos mediante un diagrama de Venn.
b) ¿Cuántos estudiantes no practican ningún deporte?
c) ¿Cuántos estudiantes juegan sólo futbol?

3. Se preguntó a 50 visitantes de un complejo de cine cuál género preferían: drama, terror o romance; 23 prefirieron drama, 18 seleccionaron terror, 13 eligieron romance, 3 optaron por drama y terror, 6 se inclinaron por drama y romance, y sólo 1 gusta de los 3 géneros.

a) Represente los subconjuntos mediante un diagrama de Venn.
b) ¿Cuántos no prefirieron ninguno de los 3?
c) ¿Cuántos prefirieron sólo drama?, ¿cuántos sólo eligieron terror? y ¿cuántos sólo seleccionaron romance?

4. Se preguntó a 200 turistas que visitaron la ciudad de México si asistieron a alguno de los 2 principales parques de diversiones de la ciudad. Se encontró que 120 visitaron el parque Los Juegos, 100 fueron a Las Atracciones y 60 visitaron los 2 parques.

a) Represente los subconjuntos mediante un diagrama de Venn e indique cuántos turistas visitaron:
i) Sólo el parque Los Juegos.
ii) Sólo Las Atracciones.
iii) Los Juegos o Las Atracciones.

5. Se preguntó a 100 estudiantes qué medio de transporte utilizan para llegar de su casa a la escuela. Se encontró que 30 lo hacen en auto, 45 en microbús, 15 en metro, 7 en auto y microbús, 12 en microbús y metro, 3 en auto y metro, mientras que 2 utilizan los 3 medios de transporte.

a) Represente los subconjuntos mediante un diagrama de Venn y señale cuántos estudiantes:
i) Llegan a la escuela sólo en auto.
ii) Utilizan sólo microbús.
iii) Utilizan sólo metro.
iv) Utilizan sólo auto o microbús.
v) Utilizan sólo auto o metro.
vi) Utilizan sólo metro o microbús.

6. En la siguiente tabla se muestran las características en cuanto a sexo y edad de 200 clientes de una tienda. Indique cuántos clientes:

a) Son hombres.
b) Son menores de 30 años.
c) Son mayores de 30 años y mujeres.
d) Son mayores de 30 y mujeres, o mayores de 30 y hombres.

Edad	Hombre	Mujer	Total
Menor de 30	60	50	110
Mayor de 30	80	10	90
Total	140	60	200

7. Se preguntó a 300 alumnos de secundaria si sabían utilizar una computadora; en la siguiente tabla se clasifican los resultados en cuanto a su conocimiento y sexo. Señale cuántos estudiantes:

a) Son mujeres.
b) No saben utilizar la computadora.
c) Saben utilizar la computadora y son hombres.
d) Saben utilizar la computadora y son mujeres, o no saben utilizar la computadora y son hombres.

Conocimiento	Hombres	Mujeres	Total
Sí	115	100	215
No	45	40	85
Total	160	140	300

4.2 Conceptos básicos, terminología y notación

Se acostumbra representar los eventos con letras mayúsculas; en el caso del lanzamiento de una moneda pudiera ser A y S (águila y sol), o también A y B. Se puede utilizar E con subíndices para representar eventos genéricos.

Se representa como P(E) la probabilidad de ocurrencia del evento E o, más simplemente, la probabilidad del evento E; la probabilidad de que salga águila en el lanzamiento de una moneda se podría representar como P(A), mientras que la probabilidad de que quede hacia arriba la cara del 4 en el lanzamiento de un dado se puede representar como P(4).

Como se verá más adelante, la probabilidad sólo puede asumir valores entre 0 y 1 (tanto por uno) o entre 0 y 100 (tanto por ciento) y se diría, por ejemplo, que la probabilidad de que salga águila en el lanzamiento de una moneda es de 0.5, o P(A)= 0.5 aunque, por otro lado, la manera en la que se expresa la probabilidad en el lenguaje cotidiano es en forma de porcentaje: "la probabilidad de que caiga águila es de 50%". Este porcentaje se calcula multiplicando la probabilidad estadística por 100 o 0.50 (100) = 50 por ciento.

4.2.1 Conceptos importantes

En seguida se presentan algunos conceptos importantes:

Experimento aleatorio. Situaciones o ensayos que implican resultados inciertos.

- **Experimento aleatorio.** Son situaciones o ensayos que implican resultados inciertos. Por ejemplo, son experimentos aleatorios lanzar una moneda al aire o un dado, porque el resultado, ya sea que salga "águila" o que caiga hacia arriba la cara del dado que tiene 5 puntos, es un resultado aleatorio. Analizar productos que salen de una línea de producción para revisar si están defectuosos o no es también un experimento aleatorio porque el resultado también lo es. Es importante observar que se considera que una situación es un experimento aleatorio porque se desea analizar el caso desde el punto de vista proba-

bilístico, y no porque se organice la situación con el propósito específico de que sea aleatoria, aunque en ocasiones sí es intencional.

Ejemplo 4.6

Son experimentos aleatorios:

- El lanzamiento de una moneda al aire.
- Los juegos de azar como la ruleta, el póker, el Melate, los pronósticos deportivos, la Lotería Nacional, etcétera.
- La extracción al azar de piezas de una línea de producción para determinar: si están defectuosas o no, su longitud en centímetros, su peso en gramos, etcétera.
- Entrevistar personas al azar para determinar: su filiación política, sus ingresos mensuales en pesos, si fuman y cuál marca de cigarros prefieren si es que lo hacen, etcétera.

- **Espacio muestral.** Se define como el conjunto de todos los resultados posibles de un experimento aleatorio: el espacio muestral del experimento aleatorio del lanzamiento de una moneda son los eventos "sol y águila". El espacio muestral del experimento aleatorio de lanzar un dado es el conjunto {1, 2, 3, 4, 5, 6,}, que son los números de puntos que tienen las 6 caras de un dado. El espacio muestral del experimento aleatorio que consiste en observar el comportamiento de los precios de las acciones que se cotizan en la Bolsa Mexicana de Valores podría ser {sube, baja, permanece igual}, o también podría ser un conjunto con un enorme número de elementos que serían los valores posibles que podría asumir la acción que se analiza.

Espacio muestral. Conjunto de todos los resultados posibles de un experimento aleatorio.

Observe que es diferente el espacio muestral, es decir, el conjunto de todos los resultados posibles, del número de elementos que hay en dicho espacio. En el caso del lanzamiento de la moneda, el número de resultados posibles es 2, y los resultados posibles son águila y sol. En el caso del dado, el número de elementos del espacio muestral es 6 y los resultados posibles son 1, 2, 3, 4, 5 y 6.

Ejemplo 4.7

Los espacios muestrales de algunos experimentos aleatorios mencionados en el ejemplo anterior son:

Experimento aleatorio	Espacio muestral
Lanzamiento de una moneda	{sol, águila}
Melate	{x/x es una de las 18 009 460 combinaciones posibles de 6 de los números del 1 al 56}
Proceso de producción	{defectuoso, no defectuoso}
	{x/x es una de la infinita cantidad de mediciones posibles en un intervalo dado de centímetros}
	{x/x es uno de los muy numerosos sueldos mensuales posibles}

Experimento aleatorio	Espacio muestral
Encuestas	{x/x es un partido político}
	{fuma, no fuma}
	{x/x es una marca de cigarros}

- **Eventos.** Son los resultados posibles del experimento aleatorio. Sol o águila son los posibles resultados del experimento aleatorio que consiste en lanzar una moneda. Que el precio suba o baje, o que permanezca igual, son los resultados posibles, es decir, los eventos posibles, del experimento aleatorio que consiste en observar el comportamiento del precio de alguna acción en el mercado de valores.

Eventos. Resultados posibles del experimento aleatorio.

EJEMPLO 4.8

Los eventos posibles en los experimentos aleatorios mencionados en el ejemplo 4.1 podrían ser:

Experimento aleatorio	Eventos posibles
Lanzamiento de una moneda	Sol y águila (o cara y cruz)
Melate	Cualquier combinación de 6 números del 1 al 56
Selección de artículos en una línea de producción	• Defectuoso o no • Longitud • Peso
Preguntas a personas	• Partido A, partido B, partido C, etcétera • $10 000, $20 000 o cualquier otra cantidad de ingresos mensuales • Peso en gramos • Fuma o no fuma

Otra manera de definir un evento es como un subconjunto del espacio muestral. En el ejemplo del lanzamiento del dado, el evento {1} es un subconjunto del espacio muestral, como también lo son {2, 4, 6} y {1, 2, 3}.

Evento simple. Se especifica de acuerdo con una sola característica. **Evento compuesto.** Está formado por 2 o más eventos simples.

- **Eventos simples y eventos compuestos.** Un **evento simple** se especifica de acuerdo con una sola característica; por ejemplo, los eventos simples del lanzamiento de un dado son los números 1, 2, 3, 4, 5 y 6. Un **evento compuesto** está formado por 2 o más eventos simples. Mientras que un evento compuesto en el experimento aleatorio de lanzar un dado podría ser la ocurrencia de un número impar, ya que incluiría a los eventos simples 1, 2 y 3.

EJEMPLO 4.9

Un experimento aleatorio consiste en lanzar 2 dados al aire.

a) El espacio muestral de este experimento aleatorio es el conjunto de los 36 pares formados por los números 1 al 6: (1, 1), (1, 2)… (6, 6), en donde el primer valor representa la cara del dado que quedó hacia arriba y el segundo valor representa el valor correspondiente al segundo dado. Este mismo espacio muestral se puede representar gráficamente como en la figura 4.2.

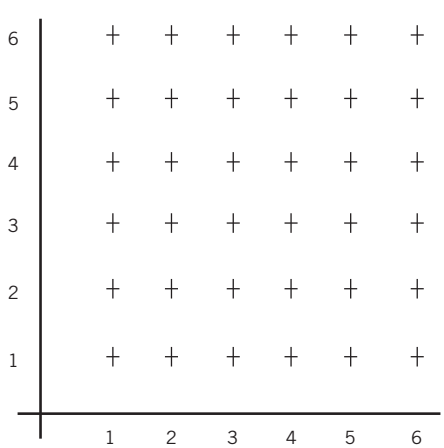


Figura 4.2 Espacio muestral para el lanzamiento de 2 dados al aire.

b) En la figura, cada punto representa un par ordenado y corresponde a uno de los eventos simples posibles en este experimento.

c) Cualquier conjunto de 2 o más de estos eventos simples sería un evento compuesto. Por ejemplo, un evento compuesto sería el formado por los posibles resultados en los que los valores de ambos dados sean iguales, o sea, (1, 1), (2, 2), (3, 3), (4, 4), (5, 5) y (6, 6). Otro ejemplo sería el conjunto de resultados para los cuales la suma de los 2 valores sea menor de 5, o sea, (1, 1), (1, 2), (2, 1), (2, 2), (3, 1) y (1, 3).

Estos 2 últimos ejemplos de eventos se ilustran en la figura 4.3.

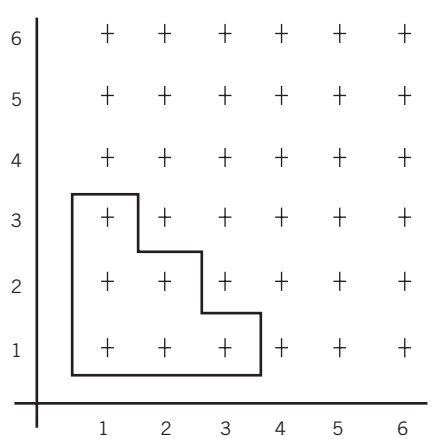


Figura 4.3 Ejemplos de eventos compuestos.

Ejemplo 4.10

Se clasificaron los 170 alumnos de un grupo escolar de acuerdo con 2 criterios: 1) estado civil (casados o solteros) y 2) sexo (hombre o mujer).

	Casados	Solteros	Total
Hombres	20	60	80
Mujeres	30	60	90
Total	50	120	170

Como se recordará, ésta es una tabla de contingencias, ya que representa datos clasificados conforme a 2 criterios y muestra, además, la información cruzada. Un experimento aleatorio consistiría en extraer de esta población un elemento al azar.

a) Si sólo se mide si la persona es casada o soltera, entonces el espacio muestral se reduce a 2 eventos posibles. Además, como son 2 los resultados posibles, el único evento compuesto del que se puede hablar es de la única combinación posible de los 2 eventos simples que abarca la totalidad del espacio muestral. Así, se podría investigar la probabilidad de que la persona extraída al azar sea casada o soltera, y este evento compuesto tiene una probabilidad de 100%, como se verá más adelante, aunque el lector pudiera saber desde este momento por qué.

b) Si se toman en cuenta los 2 criterios de clasificación, entonces el espacio muestral se forma por 4 eventos simples: 1) mujer casada, 2) mujer soltera, 3) hombre casado y 4) hombre soltero; se podría decir que el evento "casado" es un evento compuesto formado por los eventos simples "mujer casada" y "hombre casado".

En este último ejemplo se aprecia con claridad que los eventos simples y los compuestos dependen en gran medida de las características específicas del conjunto que se considera y de la forma en que se aborda el experimento aleatorio que se analiza.

Ejemplo 4.11

En el experimento de elegir una carta de una baraja inglesa, defina el espacio muestral y los eventos posibles.

Solución:

Espacio muestral	Las 52 cartas que conforman la baraja.
Eventos posibles	Cualesquiera de los números del 1 al 10, jota, reina o rey; de los 4 palos de la baraja.

Ejemplo 4.12

En el experimento de sacar 2 pelotas de un recipiente en el que se encuentran 4 de diferentes colores: azul, rojo, amarillo y naranja. Defina espacio muestral y eventos posibles.

Solución:

Espacio muestral	Las 6 combinaciones posibles entre las 4 pelotas de color.
Eventos posibles	(azul, rojo), (azul, amarillo), (azul, naranja), (rojo, amarillo), (rojo, naranja), (amarillo, naranja).

Ejemplo 4.13

En la siguiente tabla se muestran 350 clientes de un banco, clasificados según el tipo de servicio contratado y sexo. Defina el espacio muestral y los posibles eventos.

Servicio	Hombres	Mujeres	Total
Tarjeta de crédito	105	72	177
Tarjeta de débito	67	34	101
Chequera	58	14	72
Total	230	120	350

Solución:

Espacio muestral	Los 350 clientes del banco
Eventos posibles	(H, crédito), (H, débito), (H, chequera), (M, crédito), (M, débito), (M, chequera)

ejercicios 4.2 Conceptos básicos, terminología y notación

1. En el experimento de sacar una pelota de un recipiente en el que se encuentran 4 bolas de diferentes colores: azul, rojo, amarillo y naranja. Defina el espacio muestral y los eventos posibles.
2. En el experimento de escoger 2 libros al azar entre 8 publicaciones de diferentes materias: matemáticas, biología, física, química, español, civismo, historia e informática. Precise el espacio muestral y los eventos posibles si el primer libro que se toma es de matemáticas y el segundo de cualquier otra materia.
3. En la siguiente tabla se muestran las características en cuanto a sexo y edad de 200 clientes de una tienda. Defina el espacio muestral y los eventos posibles al tomar una persona al azar, si sólo se mide si es mayor o menor de 30 años.

Edad	Hombre	Mujer	Total
Menor de 30	60	50	110
Mayor de 30	80	10	90
Total	140	60	200

4. En la siguiente tabla se presentan los resultados de 500 entrevistas realizadas a compradores de valores, clasificados por sexo y tipo de valores adquiridos. Determine el espacio muestral y los eventos posibles al tomar una persona al azar, si se mide sólo si compró valores de especulación o de inversión.

Tipo de valores	Hombres	Mujeres	Total
De especulación	150	250	400
De inversión	50	50	100
Total	200	300	500

5. Se preguntó a 300 alumnos de secundaria si sabían o no utilizar una computadora. En la siguiente tabla se clasifican en cuanto a su conocimiento y sexo. Defina el espacio muestral y los eventos posibles si se toma un hombre al azar.

Conocimiento	Hombres	Mujeres	Total
Sí	115	100	215
No	45	40	85
Total	160	140	300

6. Se preguntó a 220 personas en cuál de los 4 principales medios de trasporte (auto, autobús, avión, tren) preferían viajar. En la siguiente tabla se clasifican por preferencia y si han viajado en él o no últimamente. Establezca el espacio muestral y los eventos posibles.

Medio de transporte	Sí han viajado	No han viajado	Total
Auto	48	16	64
Autobús	30	18	48
Avión	21	50	71
Tren	15	22	37
Total	114	106	220

7. En una encuesta se preguntó a 400 personas, de entre 20 y 25 años, cuál de los 7 canales principales de la televisión abierta veían más (1, 3, 6, 8, 10, 12, 14), así como en qué horario preferían sintonizarlo. En la siguiente tabla se muestran los resultados. Defina el espacio muestral así como los eventos posibles.

Canal	Matutino	Vespertino	Nocturno	Total
1	39	12	58	109
3	11	8	32	51
6	6	5	26	37
8	10	13	24	47
10	9	2	18	29
12	12	10	20	42
14	28	15	42	85
Total	115	65	220	400

4.3 Técnicas de conteo, permutaciones y combinaciones

Como el estudio de la probabilidad se basa en muchos casos en el número de resultados posibles, es necesario analizar previamente las diferentes formas en las que se determina este número de casos posibles. Esto se realiza en el análisis combinatorio, que revisa diversas reglas de conteo y los importantes conceptos de permutaciones y combinaciones:

Regla 1. Resultados posibles con número constante de eventos

Si se realiza una cantidad n de ensayos y en cada uno de ellos puede ocurrir cualquiera de k eventos diferentes, mutuamente excluyentes y colectivamente exhaustivos, entonces el número de resultados posibles es igual a:

$$k^n \tag{4.1}$$

Por ejemplo, en el lanzamiento de una moneda al aire pueden ocurrir 2 eventos diferentes: sol o águila. Entonces, si se lanza una moneda al aire 10 veces (10 ensayos), el número de resultados posibles es:

$$k^n = 2^{10} = 1\,024$$

También, al lanzar un dado, pueden ocurrir 6 eventos diferentes, por lo que si se lanza un dado 5 veces, el número total de resultados posibles es:

$$k^n = 6^5 = 7\,776$$

Una manera de considerar los casos anteriores es notar que la cantidad de eventos posibles en cada ensayo permanece constante: en el lanzamiento de la moneda, el número de resultados posibles es siempre 2 y en el caso del dado, son 6.

Además, esos 2 resultados de arrojar la moneda son mutuamente excluyentes porque no pueden ocurrir simultáneamente: ocurre uno o el otro, pero no los 2 a la vez. También son colectivamente exhaustivos porque en conjunto, sol y águila (o cara y cruz) agotan todas las posibles ocurrencias; en el lanzamiento de una moneda sólo puede ocurrir que caiga hacia arriba sol o águila. Por razones similares, los 6 eventos posibles del lanzamiento de un dado son también mutuamente excluyentes y colectivamente exhaustivos.

Regla 2. Principio de multiplicación

Si existen k_1 maneras en que se puede realizar un evento 1, k_2 maneras en que puede suceder un evento 2, k_3 maneras en que puede ocurrir un evento 3, y así sucesivamente hasta k_n maneras en que puede acontecer un evento n, entonces esos n eventos diferentes se pueden acomodar de la siguiente manera:

$$k_1 \cdot k_2 \cdot k_3 \cdot \cdots \cdot k_n \qquad \textbf{(4.2)}$$

Por ejemplo, las placas de automóviles constan de 3 números y 3 letras; la cantidad total de autos diferentes que se podrían emplacar son:

$$26 \cdot 26 \cdot 26 \cdot 10 \cdot 10 \cdot 10 = 17\,576\,000$$

Una baraja inglesa tiene 52 naipes. Por ello, el número de resultados posibles, si se extraen 5 cartas sucesivamente, es:

$$52 \cdot 51 \cdot 50 \cdot 49 \cdot 48 = 311\,875\,200$$

Esta operación es así porque, para elegir la primera carta, se tiene el total de los 52 naipes. Una vez que se saca la primera carta, para elegir la segunda ya sólo hay 51 disponibles y así sucesivamente hasta encontrar la quinta carta.

Regla 3. Número de formas en las que se pueden ordenar *n* objetos

El número de formas en que se pueden ordenar n objetos sin repetición es:

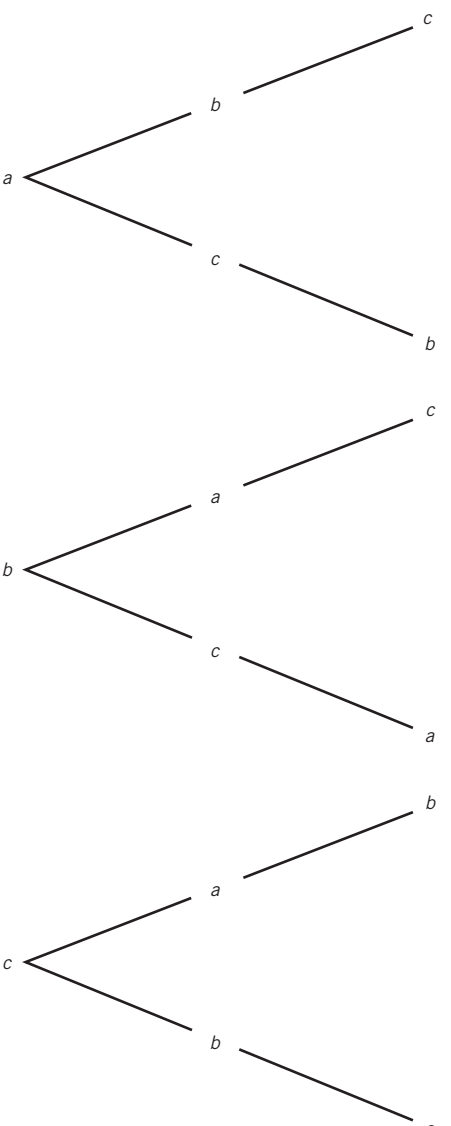
$$n! = n\,(n-1)(n-2)\,\ldots\,1 \qquad \textbf{(4.3)}$$

Por ejemplo, ¿en cuántas formas se pueden acomodar 3 personas en los lugares delanteros de un autobús?

$$3 \cdot 2 \cdot 1 = 6$$

Para situar una persona en el primer asiento se puede escoger entre cualquiera de los 3 pasajeros. Una vez asignado este primer lugar, para asignar el segundo asiento ya sólo se puede escoger entre los 2 pasajeros restantes y para asignar el último asiento ya sólo queda una persona por acomodar.

Este ejemplo conduce a otro tema importante relacionado con las técnicas de conteo y con la probabilidad: los diagramas de árbol. La situación anterior se puede ilustrar con un diagrama de árbol. Si se usan las letras a, b y c para representar a los 3 pasajeros, el diagrama se muestra a la derecha.



Permutaciones. Son todos los subconjuntos de *x* elementos que se pueden formar de entre un conjunto de *n* objetos y en donde una permutación con los mismos elementos que otra, pero en diferente orden, constituye una permutación distinta.

Estos diagramas de árbol suelen ser útiles para visualizar con mayor facilidad situaciones en las que se contemplan arreglos como el del ejemplo anterior.

Regla 4. Permutaciones

Las **permutaciones** son todos los subconjuntos de x elementos que se pueden formar de entre un conjunto de n objetos y en donde una permutación con los mismos elementos que otra, pero en diferente orden, constituye una permutación distinta.

EJEMPLO 4.14

Determinar las permutaciones de dos elementos, es decir cuando $x = 2$, del conjunto A = {azul, verde, rojo}.
Las permutaciones de tamaño 2 son:

{azul, verde}	{rojo, azul}
{verde, azul}	{verde, rojo}
{azul, rojo}	{rojo, verde}

Se trata de 6 permutaciones diferentes porque, aunque las permutaciones {azul, verde} y {verde, azul} tienen los mismos elementos, éstas no están en el mismo orden. Se puede calcular el número de permutaciones como:

$$P_n^x = \frac{n!}{(n-x)!} \tag{4.4}$$

En el ejemplo anterior, $n = 3$ y $x = 2$, por lo que:

$$P_n^x = \frac{n!}{(n-x)!} = \frac{3!}{(3-2)!} = \frac{3 \cdot 2}{1} = 6$$

Cuando $n = x$ se tiene el caso de permutaciones de n elementos tomados todos a la vez, con lo cual la fórmula se convierte en:

$$P_n^n = \frac{n!}{(n-n)!} = \frac{n!}{0!} = \frac{n!}{1!} = n!$$

Combinaciones. Son todos los subconjuntos de *x* elementos que se pueden formar de entre un conjunto de *n* objetos y en donde una combinación con los mismos elementos no es otra sino que es la misma, aunque los elementos se encuentren en diferente orden.

Regla 5. Combinaciones

Las **combinaciones** son todos los subconjuntos de x elementos que se pueden formar de entre un conjunto de n objetos y en donde una combinación con los mismos elementos no es otra combinación sino que es la misma, aunque los elementos se encuentren en diferente orden.

EJEMPLO 4.15

Determine las combinaciones de $x = 2$ del mismo conjunto A = {azul, verde, rojo} del ejemplo 4.8 anterior. Las combinaciones de tamaño 2 son sólo 3: {azul, verde}, {azul, rojo} y {verde, rojo}. Son sólo 3 combinaciones diferentes porque, por ejemplo, {azul, verde} y {verde, azul} son una sola combinación ya que tienen los mismos elementos y no importa que estén en orden diferente. Se puede calcular el número de combinaciones como:

$$C_n^x = \frac{n!}{x!(n-x)!} \tag{4.5}$$

En el ejemplo: $C_n^x = \dfrac{n!}{x!(n-x)!} = \dfrac{3!}{2!(3-2)!} = \dfrac{3 \cdot 2}{2} = 3$

EJEMPLO 4.16

Si se extrae 3 veces una carta al azar de una baraja inglesa y se reemplaza cada vez, ¿cuál es el número total de resultados posibles?

Solución: $k^n = 52^3 = 140\,608$

EJEMPLO 4.17

¿De cuántas formas se puede conformar un comité de 4 personas, de entre un grupo de 40 alumnos?

Solución: $40 \cdot 39 \cdot 38 \cdot 37 = 2\,193\,360$

Ejemplo 4.18

Con las letras A, B, C, ¿cuántas permutaciones de 3 elementos se pueden hacer?, ¿cuáles son?

Solución: $P_n^x = \frac{n!}{(n-x)!} = \frac{3!}{(3-3)!} = \frac{3 \cdot 2}{0!} = 6$

ABC	BCA
ACB	CAB
BAC	CBA

Ejemplo 4.19

En un concurso de dibujo se inscribieron 10 trabajos de los cuales se elegirán los 3 mejores. El primero obtiene una medalla de oro, el segundo de plata y el tercero de bronce. ¿Qué tipo de análisis combinatorio debe usarse?, ¿cuántas formas tienen los jueces de premiar con 3 lugares los 10 dibujos?

Solución: Son permutaciones, ya que el orden de selección hace diferencia, puesto que el primero recibe medalla de oro, y el segundo y tercer lugares, de plata y bronce, respectivamente.

$$P_n^x = \frac{n!}{(n-x)!} = \frac{10!}{(10-3)!} = \frac{10!}{7!} = 720$$

Ejemplo 4.20

En una biblioteca se quieren identificar las 45 diferentes clasificaciones que tienen con códigos de color; para ello se piensa combinar 2 colores tomando en cuenta que no importa el orden en que se encuentren. ¿Cuántos colores deben tomarse para crear los códigos, 8 o 10?

Solución: $C_n^x = \frac{n!}{x!(n-x)!} = \frac{8!}{2!(8-2)!} = \frac{8!}{2!6!} = 28$

$$C_n^x = \frac{n!}{x!(n-x)!} = \frac{10!}{2!(10-2)!} = \frac{10!}{2!8!} = 45$$

Deben tomarse 10 colores ya que las combinaciones posibles con 8 son insuficientes para clasificar todos los libros.

Ejemplo 4.21

La contraseña de una computadora se forma por 4 letras, que pueden ser cualquiera de las 27 que conforman el alfabeto.

a) ¿Cuántas contraseñas diferentes pueden formarse si el orden hace diferencia?
b) ¿Cuántas pueden formarse si el orden no hace diferencia?

Solución:

a) $P_n^x = \frac{n!}{(n-x)!} = \frac{27!}{(27-4)!} = \frac{27!}{23!} = 421\,200$

b) $C_n^x = \frac{n!}{x!(n-x)!} = \frac{27!}{4!(27-4)!} = \frac{27!}{4!23!} = 17\,550$

Ejercicios 4.3 Técnicas de conteo, permutaciones y combinaciones

1. En un sorteo se eligen 5 números del 0 al 9 y se pueden repetir. ¿Cuántas combinaciones posibles hay?
2. Se pide a una persona que ordene 5 refrescos con base en sus preferencias. ¿De cuántas formas diferentes pueden acomodarse los refrescos?
3. Se utilizarán 10 números del 0 al 9 para crear códigos de 4 dígitos, que identifiquen prendas de vestir de una tienda. ¿Cuántos códigos diferentes pueden obtenerse?
4. ¿Cuántos grupos de 3 estudiantes se pueden formar de un total de 10 si el orden no importa?
5. ¿Cuántas permutaciones y combinaciones de 2 se pueden hacer con las 5 vocales? Indique cuáles son
6. ¿De cuántas maneras diferentes se puede elegir, de entre 48 alumnos, un comité formado por un presidente, vicepresidente, secretario y tesorero?
7. ¿De cuántas maneras se pueden elegir los premios 1o., 2o. y 3o. de entre 10 vendedores de una empresa de seguros?
8. En una caja hay 20 focos que funcionan y 5 fundidos, ¿de cuántas maneras se pueden elegir al azar 4 focos que funcionan y 2 fundidos?
9. ¿De cuántas formas se pueden escoger 4 de 8 productos para colocarlos en un estante?

4.4 Interpretaciones de la probabilidad

Se revisan en esta sección las diferentes formas en las que se puede interpretar la probabilidad.

Interpretación teórica o **clásica de la probabilidad.** Parte de que cuando no hay razones para preferir uno de los posibles resultados o sucesos, se considera que todos tienen la misma probabilidad de ocurrir.

4.4.1 Interpretación teórica o clásica

La **interpretación teórica** o **clásica de la probabilidad** se basa en el principio de la razón insuficiente, el cual señala que cuando no hay razones para preferir uno de los posibles resultados o sucesos a cualquier otro, se considera que todos tienen la misma probabilidad de ocurrir.

Esta interpretación de la probabilidad es, entonces, la que se hace respecto a experimentos en los que se supone que todos los resultados tienen la misma probabilidad de ocurrencia, como en el caso del lanzamiento de una moneda, en donde se afirma que, si la moneda es legal (es decir, si no tiene truco), es igualmente probable que caiga de cualquiera de sus lados. Por ello, la probabilidad de que caiga sol es 1 de 2 o, con números:

$$P(S) = \frac{1}{2} = 0.5$$

Este mismo resultado, expresado en porcentaje, sería 0.5(100) = 50%, que es la forma en la que normalmente se expresa la probabilidad en el lenguaje cotidiano. Esta interpretación se expresa de manera formal como:

$$P(E) = \frac{n(E)}{n(EM)} \tag{4.6}$$

en donde $P(E)$ representa la probabilidad de ocurrencia de un evento cualquiera E, $n(E)$ es el número de formas en las que puede ocurrir el evento E, mientras que $n(EM)$ es el número de elementos del espacio muestral, es decir, el número total de resultados que se pueden obtener en la realización del experimento aleatorio correspondiente.

Otra manera de interpretar la probabilidad en este sentido es que, si un evento puede suceder de N formas igualmente probables, mutuamente excluyentes y colectivamente exhaustivas, y si el evento de interés E, puede ocurrir en n de estas formas, entonces la probabilidad de ocurrencia del evento E se puede expresar como:

$$P(E) = \frac{n}{N} \tag{4.7}$$

Así, esta definición implica tres características de las formas en las que puede suceder el evento:

1. Son igualmente probables, tal como plantea el principio de la razón insuficiente.
2. Son mutuamente excluyentes, lo cual quiere decir que la ocurrencia de cualquiera de los eventos posibles implica que no sucede ningún otro o, en otras palabras, que no pueden ocurrir dos resultados simultáneamente.
3. Son colectivamente exhaustivos, lo cual quiere decir que el conjunto de todos ellos representa la totalidad de las formas en las que puede ocurrir el evento.

EJEMPLO 4.22

Determine la probabilidad de obtener un 5 en el lanzamiento de un dado.

Solución: Como el dado tiene 6 lados y uno de ellos es el 5, entonces,

$$P(5) = \frac{1}{6}$$

Ejemplo 4.23

Calcule la probabilidad de que salga un número par en el lanzamiento de un dado.

Solución: Como 3 de las 6 caras del dado son números pares,

$$P(\text{par}) = \frac{3}{6} = 0.5$$

Ejemplo 4.24

¿Cuál es la probabilidad de que gane una persona que juega al Melate en la que el ganador es quien acierta a 6 de los primeros 56 enteros naturales, si juega con 10 combinaciones de 6 números?

Solución: Si juega con 10 combinaciones, entonces éste es el número de formas en las que puede suceder el evento de interés, es decir, que pueda ganar. El número total de formas en las que puede ocurrir el resultado del sorteo es igual al número de combinaciones de 6 números que se pueden formar con un total de 56 números enteros, o:

$$C_{51}^{6} = \frac{56 \cdot 55 \cdot 54 \cdot 53 \cdot 52 \cdot 51}{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = \frac{23\ 377\ 273\ 920}{720}$$
$$= 32\ 468\ 436$$

Entonces, la probabilidad de ganar el Melate con 10 combinaciones es:

$$P(E_{10}) = \frac{10}{32\ 468\ 436} = \frac{1}{3\ 246\ 844}$$

Es importante observar que esta interpretación de la probabilidad es *a priori*, ya que se pueden calcular estas probabilidades mediante razonamientos sin necesidad de experimentar los resultados; es decir, *antes* de realizar el experimento.

Por su parte, y como se analiza en el apartado siguiente, la probabilidad como frecuencia relativa sí requiere la realización del experimento o la recolección de los datos para poder evaluar las probabilidades. En otras palabras, las probabilidades se calculan *después* de recopilar los datos, por lo que se puede considerar que estas probabilidades son *a posteriori*.

Ejemplo 4.25

En un grupo de 30 mujeres hay 10 que son solteras. ¿Cuál es la probabilidad de que, al elegir una, sea soltera?

Solución:

$$P(S) = \frac{10}{30} = 0.3$$

Ejemplo 4.26

Se selecciona una carta de una baraja inglesa (52 cartas).

a) ¿Cuál es la probabilidad de que la carta elegida sea de corazones?
b) ¿Cuál es la probabilidad de que sea un rey?
c) ¿Cuál es la probabilidad de que sea un rey de corazones?

Solución:

a) $P(\text{corazones}) = \frac{13}{52} = 0.25$

b) $P(\text{rey}) = \frac{4}{52} = 0.08$

c) $P(\text{rey de corazones}) = \frac{1}{52} = 0.02$

Ejemplo 4.27

En una bolsa hay 5 canicas rojas, 10 azules, 15 negras, 8 naranjas y 2 amarillas. Si se toma una canica al azar:

a) ¿Cuál es la probabilidad de que la canica elegida sea amarilla?
b) ¿Cuál es la probabilidad de que no salga una canica negra?

Solución:

a) $P(\text{amarilla}) = \frac{2}{40} = 0.05$

b) $P(\text{no negra}) = \frac{5+10+8+2}{40} = \frac{25}{40} = 0.62$

4.4.2 La probabilidad como frecuencia relativa

En el capítulo 2 se revisaron, entre otros temas, las series simples de datos y su conversión en series de datos y frecuencias, así como las series de clases y frecuencias, y se habló también de las frecuencias relativas en la sección 2.2.3. Se revisan en seguida un par de ejemplos.

EJEMPLO 4.28

En la tabla 4.1 se resumen los resultados que se obtuvieron al analizar un embarque de 1 000 cajas de 50 focos cada una. Los datos son el número de focos defectuosos que se encontraron en cada una de las mil cajas.

Tabla 4.1. Número de focos defectuosos en 1 000 cajas de 50 focos, agrupados en una serie de datos y frecuencias, incluyendo tanto las frecuencias absolutas como las frecuencias relativas, en tanto por uno y en tanto por ciento

Número de focos defectuosos	Número de cajas (frecuencia absoluta) f	Frecuencia relativa $f_r = \frac{f}{n} = \frac{f}{1\,000}$	Frecuencia porcentual
0	700	0.70	70
1	150	0.15	15
2	100	0.10	10
3	30	0.03	3
4	10	0.01	1
5	10	0.01	1
Total	1 000	1	100

Frecuencia relativa. Es la proporción de casos en cada categoría.

La **frecuencia relativa** es, en otras palabras, la proporción de casos en cada categoría, por lo que se podría decir que 70% de las cajas de focos no contienen focos defectuosos o, abundando, que la proporción de cajas que tenía un foco defectuoso es de 15%, etcétera.

Interpretando esta situación como experimento aleatorio se diría por ejemplo que si se escoge una caja al azar, la probabilidad de que no tenga focos defectuosos es de 70%. Así, se utilizaría la frecuencia relativa como medida de la probabilidad. Esta misma interpretación de la frecuencia relativa es aplicable a tablas de contingencias o tablas de doble entrada con información cruzada.

EJEMPLO 4.29

En la tabla 4.2 se resumen los resultados de una encuesta realizada entre 1 000 ciudadanos en edad de votar y clasificadas por sexo

Tabla 4.2. Intenciones de voto de 500 ciudadanos clasificados por sexo

	Partido A	Partido B	Partido C	Partido D	Otro	Total
Hombre	150	200	120	20	10	500
Mujer	200	100	130	30	40	500
Total	350	300	250	50	50	1 000

Si se extrae de este grupo una persona al azar:

a) La probabilidad de que sea mujer es:

$$P(M) = \frac{500}{1\,000} = 0.50$$

b) La probabilidad de que el partido de su preferencia sea el partido A es:

$$P(A) = \frac{350}{1\,000} = 0.35$$

c) La probabilidad de que sea mujer y prefiera al partido B es:

$$P(M \text{ y } B) = \frac{100}{1\,000} = 0.10$$

EJEMPLO 4.30

Se contó el número de goles anotados por el equipo de futbol representante de una escuela durante los 60 partidos en todo el año.

a) ¿Cuál es la probabilidad de que no se anote gol?
b) ¿Cuál es la probabilidad de que sí se anoten goles?
c) ¿Cuál es la probabilidad de que se anoten 3 o más goles?

Goles anotados	*f*
0	13
1	22
2	10
3	8
4	5
5	2
Total	60

Solución:

Goles anotados	*f*	Frecuencia relativa	Frecuencia porcentual
0	13	0.22	22
1	22	0.37	37
2	10	0.17	17
3	8	0.13	13
4	5	0.08	8
5	2	0.03	3
Total	60	1	100

a) 0.22 o 22%
b) 0.37 + 0.17 + 0.13 + 0.08 + 0.03 = 0.78 o 78%, o también 1 – 0.22 = 0.78
c) 0.13 + 0.08 + 0.03 = 0.24 o 24%

Ejemplo 4.31

En la tabla de abajo se muestran los resultados de una encuesta realizada a 1 500 personas acerca de sus preferencias en cuanto a los principales noticieros nocturnos en los canales de televisión abierta. Se clasifican por sexo y preferencia. Si se extrae una persona al azar:

a) ¿Cuál es la probabilidad de que sea hombre?
b) ¿Cuál es la probabilidad de que prefiera el noticiero A?
c) ¿Cuál es la probabilidad de que sea mujer y prefiera el noticiero A?
d) ¿Cuál es la probabilidad de que sea hombre y prefiera el noticiero B?

Solución:

a) $P(H) = \dfrac{869}{1\,500} = 0.58$

b) $P(A) = \dfrac{606}{1\,500} = 0.404$

c) $P(M \text{ y } A) = \dfrac{238}{1\,500} = 0.16$

d) $P(H \text{ y } B) = \dfrac{138}{1\,500} = 0.09$

	Noticiero A	Noticiero B	Noticiero C	Noticiero D	Total
Hombre	368	138	40	323	869
Mujer	238	101	94	198	631
Total	606	239	134	521	1 500

4.4.3 Interpretación subjetiva de la probabilidad

Existen situaciones en las que no es posible medir la frecuencia relativa de los eventos, el número de formas en las que se puede presentar el evento de interés, el número total de eventos posibles o ninguna de las 3 cosas y, por lo tanto, no es posible aplicar la interpretación teórica o la de frecuencia relativa para calcular probabilidades; en estos casos se pueden asignar probabilidades en forma subjetiva y quien mejor puede hacer este tipo de evaluaciones es un experto en el tema.

Ejercicios 4.4 Interpretaciones de la probabilidad

1. En una caja de 25 lápices hay 2 con la punta rota. ¿Cuál es la probabilidad de que al tomar uno tenga la punta rota?, ¿cuál es la probabilidad de elegir uno que esté completo?

2. En una bodega hay 200 sillas con algún desperfecto: 40 sin respaldo, 45 sin asiento, 60 con una pata rota y 55 con 2 patas rotas. Si se toma una silla al azar:

a) ¿Cuál es la probabilidad de que no tenga respaldo?
b) ¿Cuál es la probabilidad de que no tenga ni una pata rota?

3. Se preguntó a 130 estudiantes de la licenciatura en administración de 4o. semestre cuál es el área de especializa-

ción que más les interesa. Suponiendo que se selecciona un alumno al azar:

a) ¿Cuál es la probabilidad de que le interese mercadotecnia?
b) ¿Cuál es la probabilidad de que le interese finanzas?

Área	Alumnos
Finanzas	24
Fiscal	6
Mercadotecnia	57
Operaciones	12
Recursos humanos	31

4. Se realizó un estudio para conocer el estado civil de los padres de 540 alumnos de una escuela primaria. Hay 333 parejas casadas, 183 divorciadas y 24 viudos. ¿Cuál es la probabilidad de que, al elegir un niño al azar, tenga padres divorciados?

5. Se preguntó a 2 000 conductores el número de veces que se pasaron un alto en el último mes. En la siguiente tabla se clasifican por sexo y número de veces. Si se tomara una persona al azar:

a) ¿Cuál sería la probabilidad de que no se hubiera pasado algún alto?
b) ¿Cuál es la probabilidad de que se haya pasado hasta 2 veces el alto?

	0	1	2	3	4	5 o más	Total
Hombres	880	28	14	9	5	4	940
Mujeres	1 030	18	4	3	4	1	1 060
Total	1 910	46	18	12	9	5	2 000

6. Una cadena de radio anunció que daría un premio en efectivo a 50 personas seleccionadas al azar de entre 10 000 llamadas que se recibieron durante los últimos 2 meses. ¿Cuál es la probabilidad de que los participantes ganen un premio?

7. Un pastor tiene un rebaño de 150 ovejas, si se sabe que 30 tienen parásitos, y se desparasitan 5 ovejas por día, ¿cuál es la probabilidad de que en el primer día elija a una que tenga parásitos?

Axioma. Es una proposición tan evidentemente cierta que no necesita demostración.

4.5 Axiomas de la probabilidad

Un **axioma** es una proposición tan evidentemente cierta que no necesita demostración. A continuación se explican algunos de ellos:

4.5.1 Axioma sobre los posibles valores de la probabilidad

Respecto a la probabilidad, es claro que no tiene sentido decir, por ejemplo, "la probabilidad de que llueva mañana es de 120%" porque decir que esa probabilidad es de 100% es asumir que no hay duda de que lloverá. Esto quiere decir que el valor máximo de la probabilidad, en términos porcentuales, es de 100%. Por otro lado, y por razones similares, tampoco tiene sentido hablar de probabilidades negativas, de manera que el valor mínimo de la probabilidad es 0, lo cual implica que el suceso en cuestión simplemente no tiene probabilidades de ocurrir; lo cual, siendo tan evidente que no necesita comprobación, conduce al primer axioma de la probabilidad:

$$0\% \leq P(E) \leq 100\%$$

Expresado en tanto por uno, que es como se expresa formalmente la probabilidad:

$$0 \leq P(E) \leq 1 \tag{4.8}$$

4.5.2 Axioma sobre la suma de las probabilidades de los eventos de un espacio muestral

Espacio muestral. Conjunto de todos los sucesos mutuamente excluyentes y colectivamente exhaustivos de un experimento aleatorio.

Un segundo axioma de la probabilidad es el que afirma que la suma de las probabilidades de todos los sucesos mutuamente excluyentes y colectivamente exhaustivos de un experimento aleatorio es 1. Esto se puede simplificar utilizando el concepto de **espacio muestral**, que es el conjunto de todos los sucesos mutuamente excluyentes y colectivamente exhaustivos de un experimento aleatorio, y se expresaría diciendo que la suma de las probabilidades de los eventos de un espacio muestral es 1. O, en símbolos:

$$P(E_1) + P(E_2) + \cdots + P(E_n) = 1 \tag{4.9}$$

Este mismo axioma se puede expresar de manera más compacta como:

$$\sum_{i=1}^{n} P(E_i) = 1$$

Para ilustrarlo en forma sencilla, en el experimento aleatorio que consiste en lanzar una moneda al aire se tienen 2 resultados posibles, águila y sol que, en conjunto, constituyen el espacio muestral y como la probabilidad de ocurrencia de cualquiera de ellos es de 0.5, se tiene que:

$$P(E_1) + P(E_2) = 0.5 + 0.5 = 1$$

Otro ejemplo igualmente sencillo sería el lanzamiento de un dado. En este caso el espacio muestral consta de seis sucesos posibles, la aparición de 1, 2, 3, 4, 5 o 6, cada uno igualmente probable con una probabilidad de un sexto ($\frac{1}{6}$) para cada uno de ellos. Se puede ver con facilidad que la suma de las 6 fracciones de un sexto es igual a uno:

$$\sum_{i=1}^{6} P(E_i) = 1$$

4.5.3 Axioma sobre la probabilidad de ocurrencia de dos a más eventos mutuamente excluyentes

Un tercer axioma de la probabilidad afirma que la probabilidad de que sucedan uno o más de varios eventos mutuamente excluyentes es igual a la suma de sus probabilidades. En símbolos:

$$P(A \cup B \cup C \ldots \cup X) = P(A) + P(B) + \cdots + P(X) \qquad \textbf{(4.10)}$$

Se puede ilustrar este axioma con el ejemplo del lanzamiento de un dado. ¿Cuál es la probabilidad de que aparezca un 1 o un 2 al lanzar un dado? Como se trata de eventos mutuamente excluyentes, la probabilidad es:

$$P(1 \cup 2) = P(1) + P(2) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \frac{1}{3}$$

EJEMPLO 4.32

Se preguntó a 50 personas si estaban de acuerdo con el aumento en los impuestos, de las cuales 10 respondieron que sí.

a) ¿Cuál es la probabilidad de que al elegir una persona al azar respondiera "sí"?

b) ¿Esto cumple con $0 \leq P(A) \leq 1$?

Solución: $P(\text{sí}) = \frac{10}{50} = 0.2$ sí cumple.

EJEMPLO 4.33

Se entregará un reconocimiento al mejor ensayo de un grupo de alumnos conformado por 4 hombres y 6 mujeres.

a) ¿Cuál es la probabilidad que tienen en total de ganar el reconocimiento?
b) ¿Cuál es la probabilidad de que gane un hombre?
c) ¿Cuál es la probabilidad de que gane una mujer?
d) ¿Todas estas probabilidades cumplen $0 \leq P(A) \leq 1$?
e) Los resultados de *c* y *d* ¿cumplen $P(E_1) + P(E_2) + \cdots + P(E_n) + 1$?

Solución:

a) $P(\text{ganar}) = \frac{1}{10} = 0.1$
b) $P(H \text{ y ganar}) = \frac{4}{10} = 0.4$
c) $P(M \text{ y ganar}) = \frac{6}{10} = 0.6$
d) Sí.
e) Sí: $P(H \text{ y g}) + P(M \text{ y g}) = 0.4 + 0.6 = 1$

EJEMPLO 4.34

La probabilidad de que ocurra un terremoto en México durante los próximos 15 años es de 0.8. ¿Cuál es la probabilidad de que no ocurra y por qué?

Solución: $P(\text{no terremoto}) = 0.2$

Porque $P(E_1) + P(E_2) + \cdots + P(E_n) = 1$

$P(\text{sí}) + P(\text{no}) = 0.8 + 0.2 = 1$

EJERCICIOS 4.5 Axiomas de la probabilidad

1. Un bote lleno de monedas contiene: 25 de \$1, 12 de \$2, 4 de \$5 y 1 de \$10. Si se saca una moneda al azar, ¿cuál es la probabilidad de que su denominación sea de \$1 o de \$10?
2. En un paquete hay 6 pliegos de cartulina blanca, 4 de cartulina azul y 5 de cartulina rosa. ¿Cuál es la probabilidad de que al tomar uno sea blanco, azul o rosa?
3. Un plato con rebanadas de fruta tiene: 12 de manzana, 13 de plátano, 10 de papaya, 7 de guayaba y 3 de piña. ¿Cuál es la probabilidad de que al elegir una al azar sea piña o papaya?
4. Se formó un comité de 7 miembros de una empresa para estudiar los conflictos internos de la organización. ¿Cuál es la probabilidad de que cualquiera de los 7 sea elegido vocero? ¿Cumple con $0 \leq P(A) \leq 1$?
5. En un estante se encuentran 25 revistas: 5 de ciencia, 4 de autos, 7 de salud, 6 de finanzas, 3 de sociales.

   a) ¿Cuál es la probabilidad de que al tomar una al azar, sea de:
   - *i*) ciencia,
   - *ii*) autos,
   - *iii*) salud,
   - *iv*) finanzas,
   - *v*) sociales?

   b) ¿Estas probabilidades cumplen $P(E_1) + P(E_2) + \cdots + P(E_n) = 1$?
6. Si se saca una carta de una baraja inglesa (52 cartas):

   a) ¿Cuál es la probabilidad de que sea de:
   - *i*) corazones,
   - *ii*) espadas,
   - *iii*) tréboles,
   - *iv*) diamantes?

   b) ¿Estas probabilidades cumplen $P(E_1) + P(E_2) + \cdots + P(E_n) = 1$?
7. La compañía de comida rápida Taco T tiene 245 restaurantes a lo largo de la República Mexicana; en la siguiente tabla se muestra el nombre del estado y el número de establecimientos que la cadena tiene en él. Si se toma un restaurante al azar, calcule la probabilidad de que se encuentre en cada uno de los estados e indique si cumple con $P(E_1) + P(E_2) + \cdots + P(E_n) = 1$.

Estado	Núm. de restaurantes
Distrito Federal	56
Hidalgo	53
Jalisco	43
Morelos	37
Puebla	28
Edo. de México	28
Total	245

4.6 Regla de la suma de probabilidades

La regla de la adición de probabilidades se desprende directamente del tercer axioma de la probabilidad, que se refiere a la probabilidad de ocurrencia de uno o más de entre varios eventos mutuamente excluyentes, salvo que ahora se elimina la condición de que sean eventos mutuamente excluyentes. Para revisar esta regla, conviene comenzar con un ejemplo.

EJEMPLO 4.35

En un grupo de personas, 25% fuma cigarros M, 10% fuma cigarros V y 5% fuma ambas marcas. Si se elige una persona al azar, la probabilidad de que fume cigarros M o V es:

$$P(M \cup V) = P(M) + P(V) - P(M \cap V) = 0.25 + 0.10 - 0.05 = 0.30.^*$$

Se ilustra esta situación en la figura 4.4.

Al sumar la probabilidad de que la persona fume cada una de las 2 marcas, se suma 2 veces la intersección, por lo que es necesario restarla una vez para llegar al resultado correcto. Generalizando, la regla de la suma de probabilidades para 2 eventos es:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \qquad \textbf{(4.11)}$$

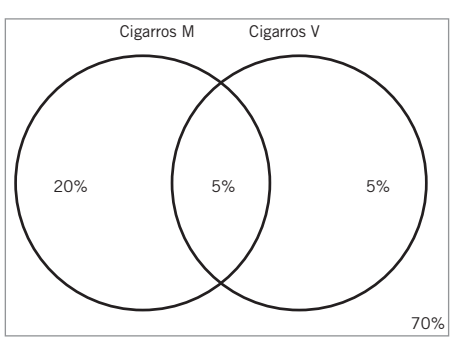


Figura 4.4 El grupo de fumadores del ejemplo 4.35.

* En este punto vale la pena notar la equivalencia entre la unión ($\cup$), la conjunción "o", la intersección ($\cap$) y la conjunción "y". Con esta equivalencia esta expresión sería: $P(M \text{ o } V) = P(M) + P(V) - P(M \text{ y } V) = 0.25 + 0.10 - 0.05 = 0.30$.

Se revisa ahora un ejemplo con eventos mutuamente excluyentes:

Ejemplo 4.36

Suponga ahora otro conjunto de personas en el que 10% fuma cigarros M y 5% fuma cigarros V, pero donde ninguna persona fuma de ambas marcas. En este caso, si se elige una persona, la probabilidad de que fume cigarros M o V es:

$$P(M \cup V) = P(M) + P(V) - P(M \cap V) = 0.10 + 0.05 - 0 = 0.15$$

Esto se ilustra en la figura 4.5.

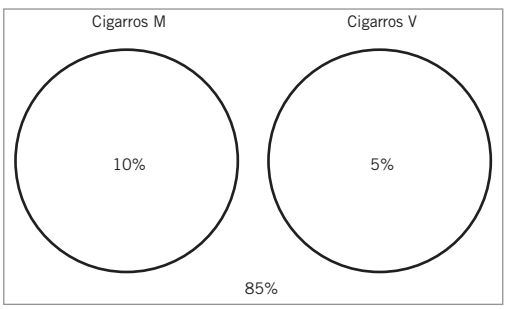


Figura 4.5 El grupo de fumadores del ejemplo 4.16.

Como se trata de conjuntos excluyentes, no hay intersección entre ellos, por lo que la probabilidad de que una persona fume de ambas marcas es 0 y, entonces, la regla de la suma de probabilidades se convierte en:

$$P(A \cup B) = P(A) + P(B) \tag{4.12}$$

La cual es la regla de la suma de probabilidades para 2 eventos mutuamente excluyentes. Resumiendo, se puede decir que $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ es la regla general de la suma de la probabilidad de 2 eventos y que $P(A \cup B) = P(A) + P(B)$ es el caso particular de la regla de la suma de probabilidades para 2 eventos mutuamente excluyentes.

Esta regla general de la suma de probabilidades se puede extender a un número n de eventos pero, entre más eventos sean, más complicada será el álgebra de conjuntos; por eso, se revisa en seguida solamente el caso de esta regla para 3 eventos.

Ejemplo 4.37

En una encuesta sobre lectura de periódicos realizada entre 1 000 personas se obtuvieron los siguientes resultados: 560 personas leen *El Trayecto* y es posible que también lean otros periódicos. 480 leen *El Global* y es posible que también lean otros periódicos, 320 leen *Cambio* y es posible que también lean otros periódicos, 140 personas leen *El Trayecto* y *El Global* y es posible que también lean *Cambio*, 150 leen *El Global* y *Cambio* y es posible que también lean *El Trayecto*, 130 leen *El Trayecto* y *Cambio* y es posible que también lean *El Global,* 60 personas leen los 3 periódicos.

En la figura 4.6 se muestra el diagrama de Venn que ilustra los resultados de esta encuesta.

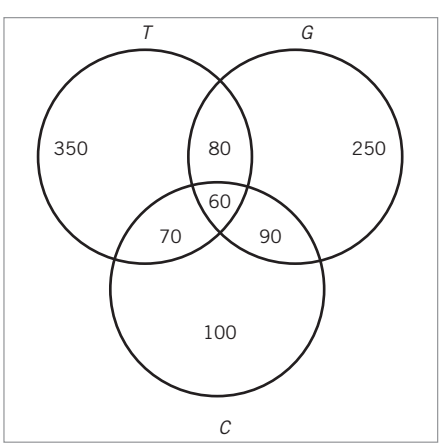


Figura 4.6 Diagrama de Venn para el ejemplo 4.17.

En primer lugar, note cómo se distribuyeron los datos en los círculos que representan a los lectores de cada periódico. Por ejemplo, para los lectores de *El Trayecto*, la suma de 350 + 80 + 60 + 70 = 560 en total, en tanto que la suma de 80 + 60 = 140 da el total de personas que leen *El Trayecto* y *El Global* y que es posible que también lean *Cambio*. El número 60 que está en la intersección de los 3 conjuntos representa al subconjunto de las personas que leen los 3 diarios.

Ahora, para ilustrar la suma del número de elementos de los 3 conjuntos, se utiliza la siguiente fórmula:

$$\begin{aligned} n(A \cup B \cup C) &= n(A) + n(B) + n(C) - n(A \cap B) \\ &\quad - n(A \cap C) - n(B \cap C) + n(A \cap B \cap C) \end{aligned} \tag{4.13}$$

la cual, utilizando las iniciales de los nombres de los periódicos para representar a los correspondientes conjuntos, tendría la forma:

$$\begin{aligned} n(T \cup G \cup C) &= n(T) + n(G) + n(C) - n(T \cap G) \\ &\quad - n(T \cap C) - n(G \cap C) + n(T \cap G \cap C) \end{aligned}$$

sustituyendo ahora los valores correspondientes se obtiene:

$$n(T \cup G \cup C) = 560 + 480 + 320 - 140 - 150 - 130 + 60 = 1\,000$$

Que es, precisamente, el número de lectores de periódicos encuestados.

Ahora, cambiando la fórmula de la suma de los elementos de 3 conjuntos a su expresión en términos de probabilidades, se tiene la regla de la suma de las probabilidades para 3 eventos que es:

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$$

Por otra parte, la regla de la suma de las probabilidades para 3 eventos mutuamente excluyentes es:

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) \qquad \textbf{(4.14)}$$

EJEMPLO 4.38

En un montón de ropa hay: 4 blusas rojas, 3 azules, 5 verdes, 1 gris y 2 negras, ¿cuál es la probabilidad de que al tomar una sea de cualquier color menos azul?

Solución: Sabiendo que la suma de las probabilidades de todos los eventos del espacio muestral es igual a uno:

$$P(A') = 1 - P(A) = 1 - \frac{3}{15} = \frac{12}{15} = \frac{4}{5} = 0.80$$

También,

$$P(r \cup v \cup g \cup n) = P(r) + P(v) + P(g) + P(n) = \frac{4}{15} + \frac{5}{15} + \frac{1}{15} + \frac{2}{5} = \frac{12}{15} = \frac{4}{5} = 0.80$$

EJEMPLO 4.39

Una máquina automática está llena de bolsas con dulces. La mayor parte de ellas contienen el peso correcto (150 g) pero algunas veces un paquete puede tener mayor o menor peso. Una revisión de 4 000 bolsas que se llenaron durante el último mes reveló lo que se muestra en la siguiente tabla. ¿Cuál es la probabilidad de que al elegir un paquete tenga más peso o que le falte?

Peso	Núm. de paquetes
Menos peso	100
Satisfactorio	3600
Más peso	300
Total	4000

Solución:

Peso	Núm. de paquetes	Frecuencia relativa
Menos peso	100	0.025
Satisfactorio	3 600	0.9
Más peso	300	0.075
Total	4 000	1

$$P(A \cup B) = P(A) + P(B) = 0.075 + 0.025 = 0.1$$

También,

$$P(A \cup B) = 1 - [P(A) + P(B)] = 1 - 0.9 = 0.1$$

EJEMPLO 4.40

¿Cuál es la probabilidad de que una carta elegida de una baraja sea un rey o un corazón?

Solución:

Carta	Probabilidad
Rey (A)	$\frac{4}{52} = 0.08$

Carta	Probabilidad
Corazones (B)	$\frac{13}{52} = 0.25$
Rey de corazones ($A \cap B$)	$\frac{1}{52} = 0.02$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.08 + 0.25 - 0.02 = 0.31$$

EJEMPLO 4.41

Se preguntó a 100 personas para qué utilizan su computadora y se encontró que: 13 la usan para hacer trabajos escolares, 53 para navegar por internet, 34 para jugar, 11 para hacer trabajos y consultar internet, 20 para navegar por internet y jugar, 1 para las 3 cosas. ¿Cuál es la probabilidad de que, al elegir a una persona, ésta la utilice para hacer trabajos, navegar por internet o jugar?

Solución:

$$p(t \cup i \cup j) = p(t) + p(i) + p(j) - p(t \cap i) - p(t \cap j) - p(i \cap j) + p(t \cap i \cap j)$$

$$p(t \cup i \cup j) = \frac{13}{100} + \frac{53}{100} + \frac{34}{100} - \frac{11}{100} - 0 - \frac{20}{100} + \frac{1}{100} = \frac{70}{100} = 0.7$$

ejercicios 4.6 Regla de la suma de probabilidades

1. En una escuela de idiomas de 300 estudiantes inscritos, 100 se encuentran en curso de inglés y 80 en curso de francés. Estas cifras incluyen a 30 inscritos en ambos cursos. Si se elige un estudiante al azar, ¿cuál es la probabilidad de que esté inscrito en inglés o en francés?
2. En las competencias de atletismo hay 120 participantes; 30 en la prueba de 100 m, 50 en la de 200 m y 40 en la carrera de relevos; pero también hay 10 que participan en las 3 pruebas. ¿Cuál es la probabilidad de que un participante esté inscrito solamente en una prueba?
3. Se evaluó a un grupo de 60 empleados de una planta ensambladora de juguetes y se encontró que 6 terminaron tarde su trabajo, 7 entregaron su trabajo con defectos y 3 entregaron tarde y con defectos. Si se elige un rebajador al azar, ¿cuál es la probabilidad de que haya entregado tarde o con defectos su trabajo?
4. En los últimos años en una empresa de servicios de telefonía se presentó una alta rotación de personal, por lo que se preguntó a los empleados que renunciaban la razón por la que se iban de la empresa: 45% dijo no estar de acuerdo con el sueldo que recibían, 55% no estaba conforme con las actividades que realizaba, 12% dijo no estar conforme con el sueldo ni con las actividades. ¿Cuál es la probabilidad de que un trabajador deje la empresa por insatisfacción con su sueldo o con las actividades que realiza?
5. Se encontró que 70% de los turistas que viajan a Italia visitan Venecia, 80% visitan Roma y 60% van a ambas ciudades. ¿Cuál es la probabilidad de que un turista vaya a Venecia o Roma?
6. Se preguntó a 1 000 personas qué bebida alcohólica preferían y se encontró lo siguiente: 729 preferían tequila, 814 mezcal y 628 vodka; 592 preferían tequila y mezcal, 465 tequila y vodka, 411 preferían mezcal y vodka y, finalmente, a 300 les gustaban las 3 bebidas. Sin embargo, al revisar los datos, se encontró que tienen un error. Indique en qué consiste ese error.
7. A 110 lectores se les preguntó cuáles secciones del periódico leían más y se encontró lo siguiente: 60 personas prefieren las noticias nacionales, 28 eligen las internacionales, 22 política, 10 nacionales e internacionales, 6 nacionales y política, 2 internacionales y política, 1 nacional, internacional, cultura y política. ¿Cuál es la probabilidad de que un lector tomado al azar lea nacionales o internacionales?
8. ¿Cuál es la probabilidad de que una carta elegida de una baraja americana sea una reina o un corazón?

4.7 Probabilidad condicional

Se utiliza la probabilidad condicional cuando la determinación de cierta probabilidad depende de circunstancias adicionales. Por ejemplo, si se tienen 2 grupos de 10 personas cada uno y el grupo A tiene 4 hombres y 6 mujeres, en tanto que el grupo B tiene 8 hombres y 2 mujeres, la probabilidad de elegir a una mujer depende del grupo que se elija; así, la probabilidad de elegir a una mujer en el grupo A es de 0.60, pues son 6 mujeres en el grupo de 10; por otra parte, la probabilidad de elegir a una mujer en el grupo B es de 0.20 porque en éste sólo hay 2 mujeres.

En términos de probabilidad condicional se dice que la probabilidad de elegir a una mujer, dado que se elige en el grupo A, es de 0.60 y, en símbolos:

$$P(M|A) = 0.60$$

La línea vertical entre la M(ujer) y A (grupo A) se lee como "dado que", por lo que ese planteamiento se puede leer como "la probabilidad de elegir a una mujer, dado que se elige en el grupo A es de 60 por ciento".

De la misma manera, la probabilidad de elegir a una mujer, dado que se elige en el grupo B es de 20% o

$$P(M|B) = 0.20$$

Se revisa en seguida otro ejemplo.

ejemplo 4.42

En el cuadro se muestran los datos de 50 vendedores que atendieron a clientes que solicitaron sus servicios para la posible compra de maquinaria especializada, divididos según si tenían experiencia en ese tipo de ventas o no y si lograron realizar la venta o no.

	Ventas logradas	Ventas fallidas	Totales
Con experiencia	15	5	20
Sin experiencia	10	20	30
Totales	25	25	50

Siguiendo el planteamiento de frecuencia relativa, si se elige al azar a cualquiera de esos vendedores, la probabilidad de que sea uno que logró la venta es de:

$$P(V) = \frac{15 + 10}{50} = 0.50$$

Si ahora se pregunta cuál es la probabilidad de que un vendedor con experiencia lograse una venta, la respuesta es:

$$P(V|E) = \frac{15}{20} = 0.75$$

La pregunta reduce la población a los 20 vendedores que tienen experiencia, por eso el planteamiento dice: "la probabilidad de que se realizó la venta (V), dado que el vendedor tenía experiencia (E) y es una probabilidad considerablemente mayor que la anterior".

Como el numerador de la expresión anterior, 15, es $n(V$ y $E)$ o, en simbología de conjuntos, $n(V \cap E) = n(E \cap V) = 15$ y el denominador es $n(E) = 20$, el planteamiento anterior puede escribirse como:

$$P(V|E) = \frac{n(E \cap V)}{n(E)} = 0.75$$

Si ahora se dividen tanto el numerador como el denominador entre el número total de vendedores $n(T)$, se tiene:

$$P(V|E) = \frac{n(E \cap V)}{n(E)} = \frac{\dfrac{n(E \cap V)}{n(T)}}{\dfrac{n(E)}{n(T)}} = \frac{P(E \cap V)}{P(E)}$$

Con lo que se expresa la probabilidad condicional $P(V|E)$ en términos de 2 probabilidades definidas en relación con la población total de vendedores. Generalizando este resultado con símbolos más comunes, se diría que si A y B son 2 eventos cualquiera y la $P(A) \neq 0$, la probabilidad condicional de B dado A, es:

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \tag{4.15}$$

Otra manera de considerar esta probabilidad condicional consiste en observar que se restringen las circunstancias originales del experimento, entonces se considera un subconjunto de la población y también puede expresarse de esta otra forma:

$$P(A|B) = \frac{P(A \text{ y } B)}{P(B)} = \frac{P(B \text{ y } A)}{P(B)}$$

EJEMPLO 4.43

Una tienda vende varios tipos de estantes en diferentes materiales y tamaños. Actualmente tiene 100 en exhibición. Si las ventas son al azar, ¿cuál es la probabilidad de que el siguiente en venderse sea de plástico, dado que ya se sabe que es de tamaño grande?

Tamaño	Plástico	Metal	Madera	Total
Chico	6	7	8	25
Mediano	23	29	2	55
Grande	12	9	4	20
Total	41	45	14	100

Solución:

$$P(\text{Grande}) = \frac{20}{100} = 0.20$$

$$P(P \text{ y } G) = \frac{12}{100} = 0.12$$

$$P(P|G) = \frac{P(P \text{ y } G)}{P(G)} = \frac{0.12}{0.20} = 0.06$$

EJEMPLO 4.44

Se observó que de los 100 niños y niñas que ingresaron a 6o. grado de primaria, sólo cierto número aprobaron el curso. ¿Cuál es la probabilidad de que al elegir un alumno sea niño aprobado?

	Niña (Na)	Niño (No)	Total
Aprobado (A)	42	29	71
Reprobado (R)	18	11	29
Total	60	40	100

Solución:

$$P(No) = \frac{40}{100} = 0.40$$

$$P(A \text{ y } No) = \frac{29}{100} = 0.29$$

$$P(A|No) = \frac{P(A \text{ y } No)}{P(No)} = \frac{0.29}{0.40} = 0.725$$

Ejemplo 4.45

En una cadena de restaurantes de comida mexicana trabajan 1 200 personas: 950 hombres y 250 mujeres. En los últimos años fueron ascendidos 324 trabajadores, según se puede ver en el cuadro siguiente:

	Hombres	Mujeres	Total
Ascendidos	288	36	324
No ascendidos	662	214	876
Total	950	250	1 200

Si se elige un trabajador al azar:

a) ¿Cuál es la probabilidad de que sea mujer dado que se sabe que es ascendido?
b) ¿Cuál es la probabilidad de que sea hombre dado que se sabe que es ascendido?

Solución:

a)
$$P(A) = \frac{324}{1\,200} = 0.27$$
$$P(M \text{ y } A) = \frac{36}{1\,200} = 0.03$$
$$P(M|A) = \frac{P(M \text{ y } A)}{P(A)} = \frac{0.03}{0.27} = 0.11$$

Nótese que esta misma probabilidad se puede calcular directamente de la tabla, sabiendo que, al decir, "dado que fue ascendido" se consideró solamente a las 324 personas que lo fueron y que, de éstas, 36 son mujeres, o sea:

$$P(M|A) = \frac{n(M \text{ y } A)}{n(A)} = \frac{36}{324} = 0.11$$

b)
$$P(A) = \frac{324}{1\,200} = 0.27$$
$$P(H \text{ y } A) = \frac{288}{1\,200} = 0.24$$
$$P(H|A) = \frac{P(H \text{ y } A)}{P(A)} = \frac{0.24}{0.27} = 0.89$$

A la igual que en el inciso anterior, directamente de la tabla:

$$P(H|A) = \frac{n(H \text{ y } A)}{n(H)} = \frac{288}{324} = 0.89$$

Ejercicios 4.7 Probabilidad condicional

1. Se tienen 2 eventos A y B; se sabe que $P(A) = 0.6$, $P(B) = 0.5$ y $P(A \text{ y } B) = 0.4$

a) Calcule la probabilidad de $P(A|B)$.
b) Calcule la probabilidad de $P(B|A)$.

2. Una discoteca tiene los siguientes datos sobre la edad y estado civil de 140 clientes.

	Soltero	Casado	Total
Menor 25	75	16	91
Mayor 25	27	22	49
	102	38	140

a) ¿Cuál es la probabilidad de que sea casado dado que se sabe que es mayor de 25?
b) ¿Cuál es la probabilidad de que sea menor de 25 dado que se sabe que es soltero?

3. A 100 asistentes a un centro comercial se les preguntó si el motivo por el que acudían a ese lugar era comodidad, variedad de tiendas o costo. En la siguiente tabla se resumen los resultados clasificados por motivo y sexo.

a) ¿Cuál es la probabilidad de que sea mujer dado que el motivo es el costo?
b) ¿Cuál es la probabilidad de que el motivo sea la comodidad dado que es hombre?

	Comodidad	Variedad	Costo	Total
Hombre	14	15	18	47
Mujer	17	24	12	53
Total	31	39	30	100

4. Una compañía elabora 2 tipos de productos: pan blanco y pan dulce. La probabilidad de que las ventas del pan blanco sean 10% en comparación con el año pasado es de 0.3, y la probabilidad de que el pan dulce se venda 10% más en comparación con el año pasado es de 0.2. La probabilidad de que ambos productos tengan ventas superiores a 10% respecto al año pasado es de 0.06. ¿Cuál es la probabilidad de que el pan blanco se venda 10% más que el año anterior, dado que el pan dulce alcanzó este punto en las ventas?

5. En un estudio entre los usuarios de aerolíneas se obtuvieron los siguientes resultados:

	Buen servicio	Mal servicio
Aerolíneas precios altos	84	36
Aerolíneas precios bajos	32	42

¿Cuál es la probabilidad de que se elija una aerolínea con buen servicio dado que es una aerolínea con precios altos?

6. Las probabilidades de que un estudiante de secundaria repruebe matemáticas, física o ambas son: $P(m) = 0.20$, $P(f) = 0.15$, $P(m y f) = 0.03$. ¿Cuál es la probabilidad de que un estudiante repruebe matemáticas dado que reprobó física?

7. Una fábrica de piezas para radios calcula que la probabilidad de que las piezas estén a tiempo para cubrir una orden es de 0.72, y que la probabilidad de que las piezas estén a tiempo y se surtan las partes a tiempo es de 0.54. ¿Cuál es la probabilidad de que las piezas estén a tiempo dado que se surte la orden a tiempo?

8. En la siguiente tabla se muestra la distribución de la población de cierta comunidad en cuanto a grupos sanguíneos.

	A	B	AB	O	
Rh+	33%	9%	4%	36%	82%
Rh–	7%	2%	1%	8%	18%
	40%	11%	5%	44%	100%

a) ¿Cuál es la probabilidad de que una persona tenga Rh– dado que tiene tipo de sangre O?

b) ¿Cuál es la probabilidad de que una persona tenga sangre tipo B dado que tiene Rh+?

9. Con base en los siguientes datos, ¿cuál es la probabilidad de que llueva dado que amaneció nublado?

a) Evento N: amanece nublado

b) Evento L: llueve

i) Amanece nublado y llueve 35% de los días.

ii) Amanece nublado y no llueve 30%.

iii) Amanece despejado y llueve 15%.

iv) Amanece despejado y no llueve 20%.

Independencia estadística. Es la probabilidad condicional de la ocurrencia de un evento si la probabilidad de la de otro no tiene efectos sobre la suya.

4.8 Independencia estadística

La **independencia estadística** se define en función de la probabilidad condicional, como sigue: se dice que 2 eventos son independientes si la ocurrencia de uno de ellos no tiene efectos sobre la probabilidad de ocurrencia del otro, lo cual puede plantearse en símbolos como:

$$P(A|B) = P(A) \quad \text{o} \quad P(B|A) = P(B) \tag{4.16}$$

Una manera sencilla de ejemplificar esto es mediante la extracción de 2 cartas de una baraja americana de 52 naipes. Si se extraen 2 cartas de una baraja y se reemplaza la primera al mazo se trata de eventos independientes porque, sin importar cuál sea la primera carta extraída, como se reemplaza al mazo, éste sigue siendo el mismo y esa primera extracción no tiene efecto alguno sobre la probabilidad de ocurrencia de la segunda extracción. Por otro lado, si se extrae la primera carta pero no se reemplaza, entonces son eventos dependientes porque el resultado de la segunda extracción depende de lo que haya ocurrido en la primera, ya que en esa segunda sólo quedarían 51 cartas y no las 52 que tiene el mazo completo.

EJEMPLO 4.46

Se sabe que en una baraja americana sólo hay un as de corazones, por lo que $P(A \text{ y } C) = \frac{1}{52}$. También se sabe que hay 13 cartas que son corazones, por lo que $P(C) = \frac{13}{52}$ y se sabe, asimismo, que hay 4 ases, con lo que $P(A) = \frac{4}{52}$. Con estos datos se puede probar si los eventos as y corazones son independientes, si se cumple que $P(A \text{ y } C) = P(A) \cdot P(C)$:

$$P(A \text{ y } C) = P(A) \cdot P(C)$$
$$\frac{1}{52} \neq \frac{4}{52} \cdot \frac{13}{52}$$

Como, evidentemente, la igualdad no se cumple, se trata, entonces, de eventos dependientes.

EJEMPLO 4.47

Calcular la probabilidad de extraer un rey de una baraja y sacar un 6 al lanzar un dado. ¿Qué tipo de eventos son?

Solución: $P(A \text{ y } B) = P(A) \cdot P(B) = \frac{4}{52} \cdot \frac{1}{6} = \frac{4}{312} = \frac{1}{78}$

Son eventos independientes ya que la carta que se saca de la baraja no influye en el lanzamiento del dado.

Sobre la independencia estadística es importante enfatizar la diferencia entre eventos independientes y eventos mutuamente excluyentes, ya que suele causar confusión. Se dice que **2**

eventos son **mutuamente excluyentes** cuando no pueden ocurrir al mismo tiempo, en tanto que **2 eventos** son **independientes** si la ocurrencia de uno de ellos no tiene efecto sobre la probabilidad de ocurrencia del otro. Se puede decir que 2 eventos mutuamente excluyentes son un caso de eventos sumamente dependientes porque la probabilidad de ocurrencia de uno de ellos necesariamente implica que la probabilidad de ocurrencia del otro es de cero.

Eventos mutuamente excluyentes. Eventos que no pueden ocurrir al mismo tiempo.
Eventos independientes. Suceden cuando la ocurrencia de uno de ellos no tiene efecto sobre la probabilidad de ocurrencia del otro.

Ejemplo 4.48

Se clasifican los empleados de una empresa de acuerdo con su sexo y su estado civil, y se tienen los siguientes resultados:

	Sexo		Totales
Estado civil	**Hombres**	**Mujeres**	
Solteros	16	24	40
Casados	24	36	60
Totales	40	60	100

¿El evento de ser soltero es independiente del evento de ser hombre?

Solución: Para verificar esto, habría que revisar si $P(S|H) = P(S)$

y, $$P(S|H) = \frac{P(S \cap H)}{P(H)} = \frac{\frac{16}{100}}{\frac{40}{100}} = 0.40$$

y, $$P(S) = \frac{40}{100} = 0.40$$

Por lo que puede decirse que, en este grupo, los 2 eventos son independientes.

Ejemplo 4.49

En una caja llena de canicas hay 10 rojas, 20 azules y 70 negras; si se extraen 4 al azar y se devuelven después de la extracción, ¿cuál es la probabilidad de que se saque una roja, una negra, una azul y otra azul?

Solución: Como son eventos independientes:

$$(R \text{ y } N \text{ y } A \text{ y } A) = P(R) \cdot P(N) \cdot P(A) \cdot P(A) = (0.1)(0.7)(0.2)(0.2) = 0.003$$

Ejercicios 4.8 Independencia estadística

1. La máquina A tiene una probabilidad de 0.3 de descomponerse en los próximos 2 años y la máquina B tiene 0.1 de probabilidad de descomponerse en los próximos 2 años. Si se considera que los eventos son independientes, ¿cuál es la probabilidad de que las 2 máquinas se descompongan al mismo tiempo?
2. En una tienda de ropa hay 3 vendedoras, cada una tiene una probabilidad de 0.25 de ausentarse. El suceso de que alguna de ellas falte es independiente de las demás. ¿Cuál es la probabilidad de que las 3 falten el mismo día?
3. Se lanzan 2 monedas. ¿Cuál es la probabilidad de que caiga cara en la primera moneda y cara en la segunda? ¿Qué tipo de eventos son?
4. Al extraer sin sustitución 2 cartas de una baraja, ¿cuál es la probabilidad de que la primera sea un as y la segunda una reina?, ¿qué tipo de eventos son?
5. En un montón de 30 playeras hay 3 tamaños diferentes: 6 chicas, 12 medianas y 12 grandes. Se van a tomar 2 al azar. ¿Cuál es la probabilidad de que ambas sean chicas?, ¿qué tipo de eventos son?

4.9 Regla de la multiplicación de probabilidades

Esta regla se desprende de la regla de la probabilidad condicional, que es:

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

Si se despeja el numerador del lado derecho se tiene que:

$$P(A \cap B) = P(B|A) \cdot P(A) \qquad \textbf{(4.17)}$$

Lo cual se conoce como la regla de la multiplicación de probabilidades y es, como se detalla más adelante, el caso general que se aplica tanto a eventos dependientes como a eventos independientes, sólo que en este último se trata de un caso particular de esta regla, el cual se analiza más adelante.

EJEMPLO 4.50

Si se recibe un envío de 1 000 focos, de los cuales 50 están defectuosos y se eligen al azar 2 de ellos, ¿cuál es la probabilidad de que ambos estén defectuosos?

Solución: Como al extraer el primero y ser defectuoso ya sólo quedan 49 defectuosos, la probabilidad es:

$$P(D_1 \text{ y } D_2) = P(D_1) \cdot P\left(D_2 \middle| D_1\right)$$

$$= \frac{50}{1\,000} \cdot \frac{49}{999} = 0.00245$$

EJEMPLO 4.51

Se sabe que, de las personas que realizan el examen de admisión al nivel medio superior, 30% tomaron un curso de preparación y que 40% de quienes tomaron el curso de preparación aprobó el examen. ¿Cuál es la probabilidad de que se elija una persona al azar entre las que presentaron el examen que haya tomado el curso y aprobado el examen?

Solución: Si C representa a las personas que tomaron el curso y A a quienes aprobaron el examen, la pregunta, en símbolos, es: $P(A \text{ y } C)$, y los cálculos:

$$P(C \text{ y } A) = P(A|C) \cdot P(C) = 0.30(0.40) = 0.12$$

4.9.1 La regla de la multiplicación para eventos independientes

La regla de multiplicación que se explicó antes es la regla general y tiene, al igual que la regla de la suma, un caso particular cuando se trata de eventos independientes. Se vio en la sección 4.8 sobre independencia estadística que, cuando se trata de eventos independientes,

$$P(A|B) = P(A)$$

Lo cual quiere decir que la ocurrencia del evento B no tiene efecto alguno sobre la probabilidad de ocurrencia de A y, por ello, esa probabilidad condicional es igual a la probabilidad simple de A y, cuando es así, la regla de la multiplicación de la probabilidad se convierte en:

$$P(A \cap B) = P(B) \cdot P(A) = P(A) \cdot P(B) \tag{4.18}$$

Que es, entonces, el caso particular de la regla de la multiplicación cuando se trata de eventos independientes.

EJEMPLO 4.52

Un envío de computadoras contiene 2 defectuosas y 98 que no tienen defectos. Si se examinan 2 computadoras al azar, ¿cuál es la probabilidad de que la primera funcione (B) y la segunda no (D) si

a) no se reemplaza la primera computadora examinada, y
b) sí se reemplaza?

Solución: *a*) Como el reemplazo de la primera computadora examinada no altera la composición del envío, las probabilidades para la extracción de la segunda computadora no se alteran, por lo que,

$$P(B \text{ y } D) = P(B) \cdot P(D) = 0.98(0.02) = 0.0196$$

b) Cuando no se reemplaza la primera computadora examinada, la probabilidad de extracción de la segunda se altera. La probabilidad de extraer una primera computadora que no tenga defectos es: $P(B) = 0.98$, en tanto que la probabilidad de extraer una segunda computadora defectuosa, dado que la primera no tenía defectos, es:

$$P(B \text{ y } D) = P(B) \cdot P\left(D \middle| B\right) = 0.98\left(\frac{2}{99}\right)$$

$$= 0.98\left(0.0202\right) = 0.0198$$

EJEMPLO 4.53

En la tabla 4.3 se muestra la composición de la población mexicana mayor de 12 años, de acuerdo con los datos del Censo de Población y Vivienda 2010, clasificada por sexo y por condición conyugal.

Tabla 4.3. Composición de la población mexicana mayor de 12 años

	Hombres	Mujeres	Total
Total	40 947 872	43 979 596	84 927 468
Solteros	15 460 577	14 392 540	29 853 117
Casados	17 067 461	17 353 462	34 420 923
En unión libre	6 045 370	6 185 310	12 230 680
Separados	970 996	2 211 430	3 182 426
Divorciados	433 354	813 202	1 246 556
Viudos	819 019	2 914 338	3 733 357
No especificado	151 095	109 314	260 409

Fuente: INEGI, disponible en: http://www3.inegi.org.mx/sistemas/TabuladosBasicos/Default.aspx?c=27302&s=est, consultado el 25 de marzo de 2011.

a) ¿Cuál es la probabilidad de elegir al azar a un mexicano (H) mayor de 12 años que sea soltero (S)?

b) ¿Cuál es la probabilidad de elegir al azar a una mexicana (M) mayor de 12 años que sea viuda (V)?

Solución: *a*)

$$P(H \cap S) = P(S) \cdot P(H|S) = \frac{29\ 853\ 117}{84\ 927\ 468} \cdot \frac{15\ 460\ 577}{29\ 853\ 117} = 0.35151(0.51789) = 0.18204$$

Pero nótese que la respuesta también se puede plantear como:

$$P(S \cap H) = P(H) \cdot P(S|H) = \frac{40\ 947\ 872}{84\ 927\ 468} \cdot \frac{15\ 460\ 577}{40\ 947\ 872} = 0.48215(0.37757) = 0.18204$$

Es fácil visualizar desde los quebrados de ambas expresiones que el resultado es el mismo. En el primer planteamiento, aparece 29 853 117 en el numerador del primer quebrado y en el denominador del segundo, por lo que la multiplicación de ambos se reduce a $\frac{15\ 460\ 577}{84\ 927\ 468}$, y lo mismo sucede en el segundo planteamiento, ya que también aparece la misma cantidad, 40 947 872, tanto en el numerador de la primera fracción (quebrado) como en el denominador del segundo, y la multiplicación se reduce a la misma, $\frac{15\ 460\ 577}{84\ 927\ 468}$.

Además, este valor se puede deducir directamente de la tabla, ya que hay 15 460 577 hombres solteros en un total de 84 927 468 mexicanos y mexicanas mayores de 12 años.

b)
$$P(M \cap V) = P(V) \cdot P(M|V) = \frac{3\ 733\ 357}{84\ 927\ 468} \cdot \frac{2\ 914\ 338}{3\ 733\ 357} = \frac{2\ 914\ 338}{84\ 927\ 468} = 0.0343$$

O, alternativamente,

$$P(V \cap M) = P(M) \cdot P(V|M) = \frac{43\ 979\ 596}{84\ 927\ 468} \cdot \frac{2\ 914\ 338}{43\ 979\ 596} = \frac{2\ 914\ 338}{84\ 927\ 468} = 0.0343$$

Además, este valor se puede deducir directamente de la tabla ya que hay 2 914 338 mujeres viudas en un total de 84 927 468 mexicanos y mexicanas mayores de 12 años.

EJERCICIOS 4.9 Regla de la multiplicación de probabilidades

En la tabla 4.3 se presenta un desglose de la población mexicana según sexo y en grupos de edades.

Tabla 4.3. Edad y sexo de la población mexicana

Edad (años)	Hombres	Mujeres	Total
0 a 10	12 146 191	11 769 700	23 915 891
11 a 20	10 927 296	10 856 730	21 784 026
21 a 30	8 938 941	9 617 150	18 556 091
31 a 40	7 898 681	8 674 176	16 572 857
41 a 50	5 943 904	6 520 204	12 464 108
51 a 60	4 043 049	4 432 217	8 475 266
61 a 70	2 405 484	2 695 674	5 101 158
71 a 80	1 313 064	1 516 824	2 829 888
81 a 90	465 190	593 386	1 058 576
91 a 99	65 984	96 812	162 796
100 o más	7 228	11 247	18 475
Total	54 155 012	56 784 120	110 939 132

Fuente: INEGI, *XII Censo General de Población y Vivienda, 2010*, disponible en: http://www3.inegi.org.mx/sistemas/TabuladosBasicos/Default.aspx?c=27302&s=est, consultado el 23 de marzo de 2011.

Con los datos de la tabla 4.3, determine:

1. La probabilidad de elegir un mexicano de 100 años o más de edad utilizando la regla de la multiplicación de probabilidades.
2. La probabilidad de elegir una mexicana de 10 años de edad o menos utilizando la regla de la multiplicación de probabilidades.
3. La probabilidad de que una persona desarrolle algún tipo de cáncer en su vida es de 0.27. Si la probabilidad de que el cáncer sea de pulmón es de 0.19, y la probabilidad de que sea leucemia es de 0.22:
   a) ¿Cuál es la probabilidad de que una persona desarrolle cáncer de pulmón?
   b) ¿Cuál es la probabilidad de que la persona desarrolle leucemia?
4. En un lote de 50 grabadoras hay 10 defectuosas. Si se eligen 2 al azar, ¿cuál es la probabilidad de que las 2 estén defectuosas?
5. En una prueba de capacidad motriz se vio que la probabilidad de que un niño de 4 años de edad realice bien todos los ejercicios es de 0.84. Si la probabilidad de que el niño que aprueba vaya a la escuela es de 0.75; y la probabilidad de que el niño apruebe pero no vaya a la escuela es de 0.25,
   a) ¿Cuál es la probabilidad de que al realizarle la prueba al niño apruebe y vaya a la escuela?
   b) ¿Cuál es la probabilidad de que apruebe y no vaya a la escuela?

4.10 Regla de Bayes

Esta regla fue planteada formalmente por Pierre Simon de Laplace (matemático, astrónomo y físico francés, 1749-1827) con base en trabajos previos del reverendo Thomas Bayes; se basa en la probabilidad condicional y su origen podría remontarse a reflexiones sobre la probabilidad de que Dios exista, dados los fenómenos que podemos observar o, en símbolos, $P(D|F)$ aunque, por otro lado, si damos por sentada la existencia de Dios como muchos lo hacen, la reflexión sería en el sentido de la probabilidad de que se den los fenómenos que observamos, dada la existencia de Dios, o $P(F|D)$.

Así, la regla de Bayes trata de problemas en los que se desea encontrar la probabilidad de un suceso B, dado otro A, $P(B|A)$, cuando los datos de los que se dispone son las probabilidades condicionales inversas, es decir, la probabilidad del suceso A, dado el suceso B, $P(A|B)$.

Para ver cómo se obtiene esta regla, se puede comenzar observando que la regla de la multiplicación de probabilidades se puede plantear de 2 maneras, considerando que $P(A \text{ y } B) = P(B \text{ y } A)$:

$$P(A \text{ y } B) = P(A|B)\,P(B) \quad \text{y} \quad P(A \text{ y } B) = P(B \text{ y } A) = P(B|A)\,P(A)$$

Si se igualan las 2 ecuaciones anteriores, eliminando $P(A \text{ y } B)$, se tiene:

$$P(A|B)P(B) = P(B|A)\,P(A)$$

Si ahora se despeja, se tiene:

$$P\left(A|B\right) = \frac{P\left(B|A\right) \cdot P(A)}{P(B)}$$

que es otra manera de interpretar la probabilidad condicional. También se tiene que:

$$P(B) = P(B|A_1) \cdot P(A_1) + P(B|A_2) \cdot P(A_2) + \cdots + P(B|A_n)\,P(A_n)$$

Lo cual, básicamente, indica que la probabilidad del evento B es la suma de las probabilidades condicionales. Sustituyendo esta última expresión en la primera, se tiene la resta de Bayes:

$$P\left(A_i|B\right) = \frac{P\left(B|A_i\right) \cdot P(A_i)}{P(B)} = \frac{P\left(B|A_i\right) \cdot P(A_i)}{\sum_{i=1}^{n} P\left(B|A_i\right) \cdot P(A_i)} \tag{4.19}$$

Ejemplo 4.54

En la siguiente tabla se resumen los datos de la fabricación de un producto mediante máquinas. En la segunda columna se muestra el porcentaje con que cada máquina contribuye a la producción; en la tercera columna se muestran los porcentajes de artículos defectuosos que cada máquina produce; en la cuarta columna se encuentra este porcentaje de defectuosos expresado como la probabilidad de defectos dada cada máquina y es, simplemente, la división del número de defectuosos por máquina entre 100:

M_i Máquina	$P(M_i)$ Contribución a la producción	% de defectuosos por máquina	$P(D\|M_i)$
1	0.40	2	0.02
2	0.30	3	0.03
3	0.20	5	0.05
4	0.10	4	0.04

Utilizando M para identificar las máquinas y D, los artículos defectuosos, la fórmula de la regla de Bayes se convierte en:

$$P(M_i|D) = \frac{P(D|M_i) \cdot P(M_i)}{\sum_{i=1}^{n} P(D|M_i) \cdot P(M_i)}$$

Su desarrollo completo, dado que se tienen 4 máquinas en el ejemplo, es:

$$P(M_i|D) = \frac{P(D|M_i) \cdot P(M_i)}{P(D|M_1) \cdot P(M_1) + P(D|M_2) \cdot P(M_2) + P(D|M_3) \cdot P(M_3) + P(D|M_4) \cdot P(M_4)}$$

Por ejemplo, la probabilidad de que, habiendo elegido un artículo defectuoso, éste provenga de la máquina 1 es:

$$P(M_1|D) = \frac{0.02(0.4)}{0.02(0.40) + 0.03(0.30) + 0.05(0.20) + (0.04(0.10)}$$

$$= \frac{0.008}{0.008 + 0.009 + 0.01 + 0.004}$$

$$= \frac{0.008}{0.031} = 0.2581 = 25.81\%$$

De la misma manera, la probabilidad de que, tras elegir un artículo defectuoso, éste provenga de las 3 máquinas restantes es:

$$P(M_2|D) = \frac{0.009}{0.008 + 0.009 + 0.01 + 0.004} = \frac{0.009}{0.031} = 0.2903$$

$$P(M_3|D) = \frac{0.01}{0.008 + 0.009 + 0.01 + 0.004} = \frac{0.01}{0.031} = 0.3226$$

$$P(M_4|D) = \frac{0.04}{0.008 + 0.009 + 0.01 + 0.004} = \frac{0.04}{0.031} = 0.1290$$

Se puede observar, además, que la suma de las probabilidades de que, tras escoger un artículo defectuoso, provenga de cada una de las 4 máquinas es igual a 1:

$$0.2581 + 0.2903 + 0.3226 + 0.1290 = 1$$

ya que estas máquinas constituyen el universo de donde pudo provenir el artículo defectuoso.

Ejemplo 4.55

Se analiza a los 2 vendedores que tiene a su cargo un supervisor de una aseguradora: Carlos y Mariana; Carlos vende 75% de las pólizas y Mariana 25%. Carlos tiene quejas en 15% de las pólizas que vende, y Mariana en 20%. Si un cliente presenta una queja, ¿cuál es la probabilidad de que Carlos vendiera la póliza?

Solución:

Evento C = Carlos vende la póliza
Evento M = Mariana vende la póliza
Evento Q = Se presenta queja

$P(C) = 0.75$
$P(M) = 0.25$

$P(Q|C) = 0.15$
$P(Q|M) = 0.20$

$$P(C|Q) = \frac{P(Q|C)P(C)}{P(Q|C)P(C) + P(Q|M)P(M)}$$

$$= \frac{(0.15)(0.75)}{(0.15)(0.75) + (0.20)(0.25)} = \frac{0.1125}{0.1125 + 0.05} = \frac{0.1125}{0.1625} = 0.69$$

EJEMPLO 4.56

En la siguiente tabla se muestran los resultados de una encuesta realizada en la Facultad de Contaduría y Administración a estudiantes de estas 2 licenciaturas acerca de sus promedios. ¿Cuál es la probabilidad de que un estudiante de contabilidad tenga un promedio de 8.0 o más, con base en la información de la proporción de estudiantes con promedios de 8.0 o más?

	Contabilidad	Administración	Total
Promedio menor 8.0	27	42	69
Promedio de 8.0 o más	11	20	31
Total	38	62	100

Solución:

Evento C = Estudiante de contabilidad.
Evento P_1 = Promedio 8.0 o mayor.
Evento P_2 = Promedio menor 8.0.

$$P(P_1) = 0.31$$
$$P(P_2) = 0.69$$
$$P(C|P_1) = \frac{11}{31} = 0.3548$$
$$P(C|P_2) = \frac{27}{69} = 0.3913$$

$$P(P_1|C) = \frac{P(C|P_1) \cdot P(P_1)}{\sum_{i=1}^{n} P(C|P_i) \cdot P(P_i)} = \frac{0.3548(0.31)}{0.3548(0.31) + 0.3913(0.69)}$$

$$= \frac{0.11}{0.11 + 0.27} = \frac{0.11}{0.38} = 0.289 = \frac{11}{38}$$

EJEMPLO 4.57

El director de una compañía productora de refrescos planea la introducción de un nuevo sabor. Anteriormente, 45% de los productos lanzados al mercado tuvieron éxito y 55% no; antes de lanzar algún refresco se realiza una investigación, de la cual se expide un informe favorable o desfavorable. En el pasado, 75% de los refrescos con éxito recibieron informes favorables y 25% de los refrescos sin éxito también recibieron un informe favorable. ¿Cuál es la probabilidad de que el nuevo sabor de refresco tenga éxito si recibe un informe favorable?

Solución:

Evento E_1 = Refresco con éxito.
Evento E_2 = Refresco sin éxito.
Evento F = Informe favorable.

$$P(E_1) = 0.45$$
$$P(E_2) = 0.55$$
$$P(F|E_1) = 0.75$$
$$P(F|E_2) = 0.25$$

$$P(E|F) = \frac{P(F|E_1)P(E_1)}{\sum_{i=1}^{n} P(F|E_i) \cdot P(E_i)} = \frac{(0.75)(0.45)}{(0.75)(0.45) + (0.25)(0.55)}$$

$$= \frac{0.3375}{0.3375 + 0.1375} = \frac{0.3375}{0.475} = 0.7105$$

EJERCICIOS 4.10 Regla de Bayes

1. En una empresa tienen 3 máquinas, A, B y C, que fabrican botones iguales; se sabe que 10% de los botones producidos por la máquina A son defectuosos así como 15% de la B y 20% de la C; también se sabe que las 3 máquinas trabajan al mismo ritmo de producción. Si se toma un botón defectuoso al azar, ¿cuál es la probabilidad de que provenga de la máquina B?
2. Tenemos 2 bolsas, B_1 y B_2. La bolsa 1 tiene 8 monedas de \$1 y 2 de \$5; mientras que la bolsa 2 tiene 4 de \$1 y 6 de \$5. Se toma una bolsa al azar y se saca una moneda de ella. La probabilidad de elegir cualquiera de las 2 bolsas es de 0.5. Si la moneda que se sacó es de \$5, ¿cuál es la probabilidad de que sea de la bolsa 1?
3. Una empresa hotelera cuenta con 3 establecimientos, del número total de habitaciones, 60% se encuentran en el hotel A, 30% en el hotel B y 10% en el hotel C. Se remodeló recientemente 9% de las habitaciones del hotel A, 20% del hotel B y 6% del hotel C. Si se elige una habitación remodelada, ¿cuál es la probabilidad de que usted se hospede en el hotel B?
4. En una empresa ensambladora de electrodomésticos hay una probabilidad de 0.3 de que alguno de sus socios les encargue fabricar un nuevo producto en los próximos 5 años y de 0.7 de que no lo haga. Si se encomendara la fabricación de un nuevo producto se tendría que contratar a más obreros; si no sucediera, habría 60% de probabilidades de contratar más trabajadores por otras razones. Si se contrataron más obreros, ¿cuál es la probabilidad de que se les encomendara la fabricación de un nuevo producto?

5. Si hay inundaciones en las zonas de cultivo del estado de Tabasco en el próximo año, la probabilidad de que el plátano aumente de precio es de 90%. Pero si no se presentan inundaciones, la probabilidad de que se incremente el precio es de 40%. Se estimó que hay 60% de probabilidades de que se presenten inundaciones en el próximo año; si el precio del plátano aumenta, ¿cuál es la probabilidad de que se deba a las inundaciones en las zonas de cultivo?

4.11 Resumen

Buena parte de la simbología que proviene de la teoría de conjuntos se utiliza con gran ventaja para revisar la teoría de la probabilidad, en particular la simbología de la unión y la intersección de conjuntos. Se utiliza la notación $P(E)$ para representar la probabilidad de un evento E cualquiera.

Se revisaron conceptos como:

- Experimento aleatorio.
- Espacio muestral.
- Eventos simples y compuestos.

En el tema de análisis combinatorio se presentaron:

- Regla 1. Resultados posibles con número constante de eventos.
- Regla 2. El principio de multiplicación.
- Regla 3. Número de formas en las que se pueden acomodar n objetos.
- Regla 4: Permutaciones.
- Regla 5: Combinaciones.

Dichas reglas son mecanismos de conteo que ayudan a calcular las probabilidades que tienen diversos eventos, el cual es tema de las siguientes subsecciones.

En el tema de interpretaciones de la probabilidad se revisaron:

- La teórica o clásica.
- La probabilidad como frecuencia relativa.
- La interpretación subjetiva de la probabilidad.

En la sección siguiente se revisaron los axiomas de la probabilidad:

- El axioma sobre los posibles valores de la probabilidad.
- El axioma sobre la suma de las probabilidades de los eventos de un espacio muestral.
- El axioma sobre la probabilidad de ocurrencia de 2 a más eventos mutuamente excluyentes.

En las secciones subsecuentes se presentaron diversas circunstancias para el cálculo de probabilidades:

- La regla de la suma de probabilidades.
- Probabilidad condicional.
- Independencia estadística.
- La regla de la multiplicación de probabilidades.
- Teorema de Bayes.

Las funciones de Excel aplicadas al tema de la probabilidad son las permutaciones y combinaciones, que se revisan en la sección 4.3, y que tienen la siguiente sintaxis:

Permutaciones: PERMUTACIONES(número, tamaño)
Combinaciones: COMBINAT(número, tamaño)

donde:

Número. Es el número entero que describe el número de objetos.
Tamaño. Es el número entero que indica el número de objetos incluidos en cada permutación o combinación.

- Permutaciones

En el ejemplo 4.14 en donde se buscan las permutaciones de x = 2 del conjunto A conformado por 3 elementos A = {Azul, Verde, Rojo}. Entonces la fórmula es:

=PERMUTACIONES(3,2)

Lo que nos da como resultado 6, exactamente igual que en el texto.

- Combinaciones

La sintaxis de la función de Excel para calcular combinaciones es:

COMBINAT(número, tamaño)

donde:

Número. Es número entero que describe el número de objetos.
Tamaño. Es el número entero que indica el número de objetos incluidos en cada permutación.

En el ejemplo 4.15 se buscan las combinaciones de x = 2 del mismo conjunto A conformado por 3 elementos A = {azul, verde, rojo}. Entonces, de igual manera, se introduce en la celda A1 el número 3 (número entero que describe el número de objetos en el conjunto), y en la celda A2 el número 2 (el número de objetos incluidos en cada permutación). Ahora en cualquier celda se escribe:

=COMBINAT(A1, A2)

Lo que nos da como resultado 3, lo mismo que en el texto.

4.12 Fórmulas del capítulo

4.3 Técnicas de conteo, permutaciones y combinaciones

Regla 1. Resultados posibles con número constante de eventos.

$$k^n \tag{4.1}$$

Regla 2. El principio de multiplicación.

n eventos diferentes se pueden acomodar de

$$k_1 \cdot k_2 \cdot k_3 \cdot \cdots \cdot k_n \text{ maneras} \tag{4.2}$$

Regla 3. Número de formas en las que se pueden acomodar n objetos.

$$n! = n(n-1)(n-2)\ldots 1 \tag{4.3}$$

Regla 4. Permutaciones.

$$P_n^x = \frac{n!}{(n-x)!} \tag{4.4}$$

Regla 5. Combinaciones.

$$C_n^x = \frac{n!}{x!(n-x)!} \tag{4.5}$$

4.4 Interpretaciones de la probabilidad

4.4.1 Interpretación teórica o clásica

$$P(E) = \frac{n(E)}{n(EM)} \tag{4.6}$$

$$P(E) = \frac{n}{N} \tag{4.7}$$

4.5 Los axiomas de la probabilidad

1. El axioma sobre los posibles valores de la probabilidad.

$$0 \le P(E) \le 1 \tag{4.8}$$

2. El axioma sobre la suma de las probabilidades de los eventos de un espacio muestral.

$$P(E_1) + P(E_2) + \cdots + P(E_n) = 1 \tag{4.9}$$

3. El axioma sobre la probabilidad de ocurrencia de 2 a más eventos mutuamente excluyentes.

$$P(A \cup B \cup C \ldots \cup X) = P(A) + P(B) + \cdots + P(X) \tag{4.10}$$

4.6 La regla de la suma de probabilidades

Para 2 eventos no mutuamente excluyentes

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{4.11}$$

Para 2 eventos mutuamente excluyentes

$$P(A \cup B) = P(A) + P(B) \tag{4.12}$$

Para 3 eventos no mutuamente excluyentes

$$n(A \cup B \cup C) = n(A) + n(B) + n(C) - n(A \cap B) - n(A \cap C) - n(B \cap C) + n(A \cap B \cap C) \tag{4.13}$$

Para 3 eventos mutuamente excluyentes

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) \tag{4.14}$$

4.7 Probabilidad condicional

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \tag{4.15}$$

4.8 Independencia estadística

$$P(A|B) = P(A) \tag{4.16}$$

4.9 Regla de la multiplicación de probabilidades

$$P(A \cap B) = P(B|A) \cdot P(A) \tag{4.17}$$

La regla de la multiplicación para eventos independientes.

$$P(A \cap B) = P(B) \cdot P(A) = P(A) \cdot P(B) \tag{4.18}$$

4.10 Regla de Bayes

$$P(A_i|B) = \frac{P(B|A_i)\,P(A_i)}{P(B)} = \frac{P(B|A_i)\,P(A_i)}{\sum_{i=1}^{n} P(B|A_i)\,P(A_i)} \tag{4.19}$$

4.13 Ejercicios adicionales

4.1 Teoría de conjuntos y teoría de la probabilidad

1. Se preguntó a 80 personas cuál es la estación del año que más les gusta; 12 respondieron primavera, 33 verano, 12 otoño, 21 invierno, 4 eligieron verano y primavera, 3 otoño e invierno, y 2 las 4 estaciones.

a) Represente los subconjuntos mediante un diagrama de Venn.
b) Indique a cuántas personas:
  i) Les gusta más la primavera y el verano.
  ii) Les gusta más el invierno.
  iii) Les gustan por igual las 4 estaciones.

2. Cincuenta personas que acostumbran comer comida rápida dieron los siguientes resultados en cuanto a su preferencia por estos alimentos: 22 gustan de las hamburguesas, 17 de las pizzas, 11 de los hot dogs, 12 de las hamburguesas y las pizzas, 3 de los hot dogs y las hamburguesas.

a) Represente los subconjuntos mediante un diagrama de Venn.
b) Indique cuántas personas prefieren:
  i) Hamburguesas, hot dogs o pizzas.
  ii) Pizzas.
  iii) Pizzas y hamburguesas.

3. Se preguntó a 100 amas de casa en dónde acostumbran comprar frutas y verduras, y se obtuvieron los siguientes resultados: 42 las compran en el mercado, 32 en el supermercado, 47 en el tianguis, 12 en el mercado y en el tianguis, 8 en el mercado y en el supermercado, 4 en los 3.

a) Represente los subconjuntos mediante un diagrama de Venn.
b) Señale cuántas amas de casa compran:
i) En el supermercado.
ii) En el tianguis.
iii) En el mercado.
iv) En los 3.

4. Se realizó una encuesta entre 40 estudiantes de bachillerato acerca de las actividades extracurriculares que realizan y se encontró que 11 practican algún deporte, 23 toman clases de idiomas, 7 acuden a cursos de computación, 3 practican deporte y toman clases de idiomas.

a) Represente los subconjuntos mediante un diagrama de Venn.
b) Indique cuántos estudiantes:
i) Solamente practican deporte.
ii) Solamente toman clases de idiomas.
iii) Estudian computación.

5. Se preguntó a 200 personas qué medio de transporte creían que era el más seguro para viajar y respondieron lo siguiente: 97 creían que avión, 67 autobús, 38 auto, 11 avión y autobús, 8 autobús y auto.

a) Represente los subconjuntos mediante un diagrama de Venn.
b) Indique cuántas personas creen que es más seguro:
i) Viajar en avión.
ii) Viajar en avión y autobús.
iii) Los 3 son igual de seguros.

4.2 Conceptos básicos, terminología y notación

6. En una caja hay 3 lápices de diferentes colores: morado, rojo y azul. Si se sacan 2 al azar, defina el espacio muestral y los eventos posibles sin importar que se repitan los elementos.

7. En un grupo de personas hay 2 casados (C) y 3 solteros (S), ¿cuántos grupos de 3 podemos formar? Defina el espacio muestral y los eventos posibles.

8. En una caja con 6 vasos hay 2 rotos. Si se sacan 2 vasos al azar, ¿cuál es la probabilidad de que al menos se saque uno roto? Defina el espacio muestral y los eventos posibles.

9. ¿De cuántas maneras podemos combinar 4 etiquetas de colores (azul, rojo, morado y naranja) en grupos de 3, si no importa el orden? Defina el espacio muestral y los posibles eventos.

10. En una bolsa hay 5 chicles de diferentes sabores (limón, naranja, sandía, chile, fresa y menta) los cuales se repartirán de 2 en 2 entre varios niños. ¿De cuántas maneras diferentes se podrían repartir los chicles? Defina el espacio muestral y los posibles eventos.

4.3 Técnicas de conteo, permutaciones y combinaciones

11. En una bolsa hay 10 canicas de diferentes colores de la cual se sacan 3. ¿De cuántas maneras diferentes pueden hacerse las extracciones?

12. Se quiere crear claves numéricas para los artículos de una tienda de 4 dígitos cada uno. Si se toman los números del 0 al 9 y no se pueden repetir los dígitos dentro del código, ¿cuántos códigos diferentes podemos crear?

13. Se tienen 5 tarjetas de diferentes colores. Si se toma del mazo una carta al azar en 5 ocasiones, ¿de cuántas maneras diferentes se pueden sacar las cartas?

14. De un total de 87 declaraciones de impuestos se seleccionan 10 para revisarlas. ¿Cuántas muestras posibles hay?

15. De un grupo de 20 alumnos se seleccionarán 4 para formar un comité que represente a la escuela en un concurso de ortografía. ¿Cuántos comités diferentes pueden formarse?

16. De un total de 5 maestros de literatura y 7 de redacción se forma un grupo en donde debe haber 2 maestros de literatura y 3 de redacción. ¿De cuántas maneras diferentes se puede formar este grupo?

17. Una empresa contrata 3 nuevos operadores y los quiere colocar en alguna de sus 10 fábricas. ¿De cuántas maneras diferentes puede hacerlo?

18. ¿Cuántas permutaciones de 2 se pueden hacer con las letras a, b, c? Indique cuáles son.

19. De un grupo de 25 trabajadores se debe elegir un presidente y un secretario para el comité administrativo. ¿De cuántas maneras diferentes se pueden designar estos puestos?

4.4 Interpretaciones de la probabilidad: teórica o clásica

20. De 80 personas que solicitan un trabajo, 30 son pasantes. Si se elige a una persona al azar, ¿cuál es la probabilidad de que no sea pasante?

21. Se tienen 8 personas que hablan inglés, 4 que hablan francés, 3 que hablan italiano y 2 que hablan portugués. Se elige una persona al azar. Calcule:

a) La probabilidad de que esa persona hable italiano.
b) La probabilidad de que ese individuo no hable portugués.

4.4.2 La probabilidad como frecuencia relativa

22. En la siguiente tabla se muestran los datos del número de autos alquilados en una agencia de renta de automóviles en los últimos 100 días.

a) ¿Cuál es la probabilidad de que se renten exactamente 7 autos?
b) ¿Cuál es la probabilidad de que se renten 6 o más autos?

Autos (X)	Días
3	6
4	14
5	24
6	28

(*continúa*)

(*continuación*)

Autos (X)	Días
7	20
8	8
Total	100

23. Se contó el número de goles que el principal anotador de un equipo de futbol escolar logró meter durante los 25 partidos de la temporada pasada.

a) ¿Cuál es la probabilidad de que en un partido anote 5 goles?
b) ¿Cuál es la probabilidad de que anote hasta 3 goles?

Goles (X)	Partidos
0	3
1	9
2	7
3	4
4	1
5	1
Total	25

24. En un grupo de 2o. de secundaria se registró el número de estudiantes que no asistieron cada día durante los últimos 2 meses.

a) Calcule la probabilidad de que falten, a lo mucho, 3 alumnos.
b) ¿Cuál es la probabilidad de que falten 4 alumnos?
c) ¿Cuál es la probabilidad de que hayan faltado más de 6 alumnos?

Estudiantes ausentes (X)	Días
0	13
1	9
2	8
3	7
4	5
5	6
6	5
7	4
8	3
Total	60

25. Se preguntó a 150 conductores el número de infracciones que se cometieron durante el último mes. Si se elige una persona al azar:

a) ¿Cuál es la probabilidad de que no se cometieran infracciones?
b) ¿Cuál es la probabilidad de que se cometieran 7 infracciones?

Infracciones (X)	Núm. de conductores
0	53
1	33
2	21
3	15
4	11
5	9
6	5
7	3
Total	1 50

26. En la tabla que se muestra a continuación se registró el número de pedidos que tuvo una pizzería durante los últimos 60 días.

a) ¿Cuál es la probabilidad de que en un día tengan entre 10 y 13 pedidos?
b) ¿Cuál es la probabilidad de que se reciban exactamente 5 pedidos?
c) ¿Cuál es el número esperado de pedidos?

Pedidos (X)	Días
1	3
3	5
5	7
6	10
7	12
9	8
10	6
12	5
13	4
Total	60

27. En una tienda de electrónica se registró el número de ventas de consolas de videojuego durante los últimos 30 días. Calcule:

a) La probabilidad de que no se venda ninguna consola.
b) La probabilidad de que se vendan más de 2 consolas.

Consolas vendidas	Días
0	8
1	12
2	5
3	3
4 o más	2

4.5 Axiomas de la probabilidad

28. De los 50 alumnos de 1o. de primaria, 29 saben leer y escribir.

a) ¿Cuál es la probabilidad de que al elegir uno al azar sepa leer y escribir?
b) ¿Cumple con $0 \leq P(A) \leq 1$?

29. En un lote de 30 baterías para auto hay 5 defectuosas. Si se eligen 10 al azar:

a) ¿Cuál es la probabilidad de que haya 2 defectuosas en la muestra de 10 baterías?
b) ¿Cumple con $0 \leq P(A) \leq 1$?

30. Se le preguntó a 100 personas si creían que la inseguridad disminuyó en los últimos meses y 12 individuos respondieron que sí.

a) ¿Cuál es la probabilidad de que, al elegir a una persona, responda que sí?
b) ¿Cumple con $0 \leq P(A) \leq 1$?

31. La probabilidad de que llueva durante los meses de julio a agosto en la ciudad de México es de 0.57. ¿Cuál es la probabilidad de que no llueva y por qué?

32. De los 35 trabajadores de un departamento, 23 son casados y 12 solteros. Si se elige al azar a un trabajador:

a) ¿Cuál es la probabilidad de que sea casado?
b) ¿Cuál es la probabilidad de que sea soltero?
c) ¿Estas 2 probabilidades cumplen con $P(E_1) + P(E_2) + \cdots + P(E_n) = 1$?

33. En una caja de 50 lápices hay 19 sin goma.

a) ¿Cuál es la probabilidad de sacar un lápiz sin goma?
b) Si se le suma la probabilidad de sacar uno con goma, ¿cumple con $P(E_1) + P(E_2) + \cdots + P(E_n) = 1$?

4.6 Regla de la suma de probabilidades

34. Se tienen 20 papeles con los nombres de personas de diferentes nacionalidades: 6 mexicanos, 4 españoles, 5 ingleses, 3 argentinos y 2 japoneses. Si se toma un papel al azar, ¿cuál es la probabilidad de que la persona no sea española?

35. En un paquete de plumas de 30 piezas hay: 15 negras, 6 rojas, 5 azules y 4 verdes. Si se toma una aleatoriamente:

a) ¿Cuál es la probabilidad de que sea negra, roja o azul?
b) ¿Cuál es la probabilidad de que sea negra, roja o verde?
c) ¿Cuál es la probabilidad de que no sea negra?

36. En un bote con 50 crayolas hay: 5 completas, 12 sin punta, 8 rotas y 25 sin cubierta. ¿Cuál es la probabilidad de que, al tomar una, esté completa o con la punta rota?

37. En un grupo de 30 hombres hay: 15 solteros, 4 casados, 6 casados y con hijos, y 5 divorciados. ¿Cuál es la probabilidad de que al escoger a alguno sea soltero o divorciado?

38. Los eventos A y B son mutuamente excluyentes. Supongamos que $P(A) = 0.30$ y $P(B) = 0.15$.

a) ¿Cuál es la probabilidad de que ocurra A o B?
b) ¿Cuál es la probabilidad de que no ocurra A o B?

39. En una tómbola se sortean diferentes premios: 15 licuadoras, 8 despensas, 5 extractores de jugo y 2 lavadoras. Si una persona saca un boleto:

a) ¿Cuál es la probabilidad de que se gane una licuadora o una despensa?
b) ¿Cuál es la probabilidad de que no se gane una licuadora?

4.7 Probabilidad condicional

40. En la siguiente tabla se muestra el total de estudiantes de una universidad clasificados por sexo y rango de edad. Si se elige a una persona al azar:

a) ¿Cuál es la probabilidad de que sea hombre dado que tiene 25 años o más?
b) ¿Cuál es la probabilidad de que sea hombre dado que tiene menos de 25 años?

	Hombre	Mujer	Total
Menos de 25 años	320	402	722
25 años o más	115	163	278
Total	435	565	1 000

41. Una empresa organiza un concurso entre sus consumidores en las 2 ciudades que registran las ventas más importantes. Para que puedan participar se tiene que mandar una carta que contenga la etiqueta de alguno de los 2 productos. Se reciben 150 000 cartas, que se clasifican en la siguiente tabla. Si se elige al ganador al azar:

	Mermelada	Verduras enlatadas	Total
Ciudad 1	28 000	83 000	111 000
Ciudad 2	22 000	17 000	39 000
Total	50 000	100 000	150 000

a) ¿Cuál es la probabilidad de que gane una carta enviada de la ciudad 1 dado que mandó una etiqueta de mermelada?
b) ¿Cuál es la probabilidad de que se mande una etiqueta de verduras enlatadas dado que la carta fue mandada de la ciudad 2?

42. Una revista de negocios realizó entrevistas a los 360 líderes de las empresas más importantes alrededor del mundo, los cuales se clasificaron por rango de edad y sexo en la siguiente tabla:

	Hombre	Mujer	Total
Menos de 45 años	94	21	115
45 o más años	172	73	245
Total	266	94	360

a) ¿Cuál es la probabilidad de que al elegir un entrevistado tenga menos de 45 años dado que es mujer?
b) ¿Cuál es la probabilidad de que sea hombre dado que tiene 45 o más años?

43. Se realizó una encuesta entre jóvenes que estudian el bachillerato acerca de cuánto y para qué utilizan internet.

	Chat	Correo electrónico	Búsqueda de información	Total
Menos de 5 horas	51	43	31	125
5 horas o más	63	41	16	120
Total	114	84	47	245

a) ¿Cuál es la probabilidad de que, al elegir uno, utilice el internet 5 horas o más dado que utiliza correo electrónico?
b) ¿Cuál es la probabilidad de que lo utilice para chatear dado que navega en internet menos de 5 horas?
c) ¿Cuál es la probabilidad de que lo utilice menos de 5 horas dado que busca información?

44. La comisión de vialidad clasificó los accidentes automovilísticos de los últimos meses con base en la hora y sexo del conductor.

	Hombre	Mujer	Total
Día	47	53	100
Noche	66	34	100
Total	113	87	200

a) ¿Cuál es la probabilidad de que un accidente ocurra durante el día dado que el conductor es una mujer?
b) ¿Cuál es la probabilidad de que el conductor sea una mujer dado que el accidente ocurre de noche?

4.8 Independencia estadística

45. Una empresa lanzará 2 productos nuevos al mercado el próximo año; después de una investigación se encontró que la probabilidad de que el producto A tenga éxito es de 0.45 y la probabilidad de que el producto B tenga éxito es de 0.7; ¿cuál es la probabilidad de que ambos tengan éxito?

46. Un transformador requiere 3 engranes para funcionar. El primer engrane tiene una probabilidad de descomponerse del 0.15, el segundo de 0.3 y el tercero de 0.21; si cada uno funciona de manera independiente, ¿cuál es la probabilidad de que los 3 se descompongan al mismo tiempo?

47. Se tienen 5 diamantes, 3 rubíes, 2 perlas y 2 esmeraldas. Si se toman 3 piedras al azar y se devuelve cada una después de tomarla:

a) ¿Cuál es la probabilidad de que se seleccione un diamante, una perla y una esmeralda?
b) ¿Cuál es la probabilidad de que se elija un rubí, una perla y un diamante?

48. Una línea de transporte escolar cuenta con 4 camiones, los vehículos A, B y C tienen 0.3 de probabilidad de encontrarse en mantenimiento, y el D, 0.4:

a) ¿Cuál es la probabilidad de que todos se encuentren en mantenimiento al mismo tiempo?
b) ¿Cuál es la probabilidad de que A, C y D se encuentren en mantenimiento el mismo día?

49. Las quejas que recibe el departamento de atención a clientes de un supermercado son de 2 tipos. La probabilidad de que sea por falta de artículos es de 0.53 y la probabilidad de que sea por lentitud en el cobro es de 0.12. ¿Cuál es la probabilidad de que una queja se presente por las 2 razones?

4.9 La regla de la multiplicación de probabilidades

50. En una ensambladora de autos se reciben de un proveedor las piezas necesarias para armar el motor. Si se reciben 200 piezas y 30 de ellas son defectuosas, ¿cuál es la probabilidad de que al tomar 2 piezas al azar durante la revisión resulten defectuosas?

51. En una bolsa de canicas hay: 5 transparentes, 3 arco iris y 2 negras. Si se toman 2 al azar:

a) ¿Cuál es la probabilidad de que las 2 sean transparentes?
b) ¿Cuál es la probabilidad de que una sea transparente y la otra negra?

52. Si en una caja de 12 crayolas 3 no tienen punta, ¿cuál es la probabilidad de que, al tomar 2, una tenga punta y la otra no?

53. Un paquete de galletas contiene 8 piezas; se observa que la mayoría de los paquetes traen 4 galletas rotas. Si se toman 2 paquetes de galletas y se sacan 4 galletas, ¿cuál es la probabilidad de que las primeras 2 estén rotas?

54. En un paquete de 100 hojas de colores hay: 21 rojas, 17 negras, 12 amarillas, 23 verdes, 19 naranjas y 8 moradas. Si se sacan 3 al azar:

a) ¿Cuál es la probabilidad de que sean de color rojo, amarillo y naranja?
b) ¿Cuál es la probabilidad de que sean de color negro, verde y morado?
c) ¿Cuál es la probabilidad de que sean de color amarillo, verde y naranja?

4.10 Regla de Bayes

55. En la caja 1 se tiene una pelota roja y una pelota azul; en la caja 2 se tienen 2 pelotas rojas; se elije una de las cajas aleatoriamente y se saca una pelota al azar.

a) Si se saca una pelota roja, ¿cuál es la probabilidad de que provenga de la caja 1?
b) Si se saca una pelota azul, ¿cuál es la probabilidad de que provenga de la caja 1?

56. En una planta productora de focos se quiere saber en qué turno se fabricó un foco que no sirve. Hay 3 turnos: matutino, vespertino y nocturno. Del total de focos que se produjeron en la planta, 40% se fabricaron en el turno matutino, 35% en el vespertino y 25% en el nocturno. Cinco por ciento de los focos producidos en la mañana no funcionaron, así como 15% de la tarde y 10% del turno nocturno. ¿Cuál es la probabilidad de que el foco defectuoso se fabricara en el turno matutino?

57. En una ONG (organización no gubernamental) se planeaba presentar una demanda judicial en contra de las tarifas de seguros de automóviles en una de 3 ciudades: Buenos Aires, Montevideo y Bogotá. La probabilidad de que se escogiera Buenos Aires era de 0.40; de que se escogiera Montevideo era de 0.35 y de que se escogiera Bogotá era de 0.25. El grupo sabía, además, que tenía 60% de probabilidades de conseguir un dictamen favorable si se seleccionaba Montevideo, 45% si se seleccionaba Buenos Aires y 35% si seleccionaba Bogotá. Si el grupo obtuvo un dictamen favorable, ¿cuál de las ciudades tuvo mejores probabilidades de ser escogida para presentar la demanda?

58. Un almacén importante considera cambiar su política de otorgamiento de crédito para reducir el número de clientes (deudores) que finalmente no pagan sus cuentas. El gerente de crédito sugiere que, a futuro, el crédito se le cancele a cualquier cliente que demore una semana o más en sus pagos en 2 ocasiones distintas. La sugerencia del gerente de crédito se basa en el hecho de que, en el pasado, 90% de los clientes que finalmente no pagaron sus cuentas se demoraron en sus pagos por lo menos 2 ocasiones. Suponga que de una investigación independiente se encuentra que 2% de todos los deudores finalmente no pagan sus cuentas y que de aquellas que sí las pagan, 45% se demoró en por lo menos 2 ocasiones. Encuentre la probabilidad de que un cliente que ya se demoró por lo menos en 2 ocasiones finalmente no pague su cuenta y con la información obtenida analice la política que sugirió el gerente de ventas.

59. Cinco líneas de producción en una operación manufacturera producen un fusible electrónico. Los fusibles son caros y se envían a los distribuidores en lotes de 100 unidades. La mayoría de los compradores prueba únicamente un pequeño número de los fusibles antes de decidir si aceptan o rechazan los lotes de fusibles que llegan, ya que la prueba es destructiva. Las 5 líneas producen fusibles a la misma velocidad y normalmente producen sólo 2% de los fusibles defectuosos que se distribuyen aleatoriamente en el proceso de producción. Desafortunadamente la línea 1 sufrió una falla mecánica y produjo 5% de artículos defectuosos en marzo. El productor se enteró de esta situación después de enviar los fusibles. Un cliente adquirió un lote producido en el mes de marzo y probó 3 fusibles y uno era defectuoso. ¿Cuál es la probabilidad de que el lote haya salido de una de las otras 4 líneas?

capítulo 5

Distribuciones discretas (discontinuas) de probabilidad

Sumario

5.1. Introducción
5.2 Distribuciones de probabilidad de variables aleatorias discretas
5.3 Media y varianza de una distribución de probabilidades
5.4 Distribución binomial
5.4.1 Media y varianza de la distribución binomial
5.4.2 Distribución binomial, tablas de probabilidades binomiales y Excel
5.5 Tres formas de presentar una distribución de probabilidad
5.6 Distribución de Poisson
5.6.1 Distribución de Poisson, tablas de probabilidades Poisson y Excel
5.6.2 Distribución de Poisson como aproximación de la distribución binomial
5.6.3 Media y varianza de la distribución de Poisson
5.7 Distribución hipergeométrica
5.7.1 Media y desviación estándar de la distribución hipergeométrica
5.8 Distribución multinomial
5.8.1 Media y desviación estándar de la distribución multinomial
5.9 Resumen
5.10 Fórmulas del capítulo
5.11 Ejercicios adicionales

En este capítulo y el siguiente se revisa el tema de las distribuciones de probabilidad, dividido en distribuciones discretas, también conocidas como *discontinuas*, que es el tema del presente capítulo, y distribuciones continuas, que se tratará en el siguiente.

Variable discontinua. Resulta clara la diferencia entre un valor y el que le sigue.
Variable continua. La diferencia entre un valor y el que le sigue es indistinguible.

Como se recordará, la distinción entre variables discontinuas y las continuas gira, precisamente, en torno de la continuidad. Una **variable** es **discontinua** si resulta clara la diferencia entre un valor y el que le sigue. Un ejemplo claro de esto es el número de hijos: los valores posibles son 0, 1, 2, etc., y la diferencia entre cualquiera de ellos y el que le sigue o le antecede es clara; por otro lado, una **variable** es **continua** si esta diferencia es indistinguible como, por ejemplo, en las mediciones: la diferencia entre la longitud de una pieza que mide 1.0000000000001 y 1.0000000000002 metros es prácticamente inexistente.

Así, las distribuciones que se revisan en este capítulo se asocian a variables discretas y son las distribuciones:

- Binomial.
- Poisson.
- Hipergeométrica.
- Multinomial.

En el capítulo siguiente se revisan las distribuciones asociadas con variables continuas y son la famosa distribución normal y la exponencial.

5.1 Introducción

Como se vio antes, cuando se realizan estudios estadísticos, los datos que se obtienen son mediciones que pueden darse en diferentes escalas (nominal, ordinal, de escala o de razón). Al medir una característica de interés, el resultado puede asumir diferentes valores: se dan variaciones en las dimensiones de piezas fa-

bricadas en máquinas, en los pesos de los animales, en el sexo, ocupación e ingresos de los elementos de la población económicamente activa, etc. Es por esta razón que se habla de **variables**.

Variable. Característica que puede asumir diferentes valores.

Además, como en muchos casos los valores de la variable que se observan se obtienen mediante procedimientos al azar, se habla de *variables aleatorias*. En general, entonces se puede decir que una **variable aleatoria** es aquella cuyo valor numérico se determina mediante el resultado de una situación incierta.

Variable aleatoria. Es aquella cuyo valor numérico se determina mediante el resultado de una situación incierta.

Por otro lado, y tal como se vio en el capítulo introductorio, una **variable discreta** es aquella en la que pueden distinguirse sin lugar a dudas 2 valores contiguos: son variables discretas el número de hijos por familia, el sexo de las personas, el número de piezas defectuosas por lote, etc. Se trata por lo general de conteos; por otro lado, las **variables continuas** son aquellas cuyos valores pueden medirse con tal precisión que se llega a perder, o se vuelve insignificante, la diferencia entre uno de sus valores y el siguiente. El peso de un animal, la estatura de una persona y la distancia recorrida son ejemplos de variables continuas. En este caso se trata, por lo general, de mediciones que pueden volverse tan precisas como el instrumento de medición o la imaginación lo permitan: la diferencia entre un peso de 54 kilogramos y otro de 54.00000000001 kilogramos es prácticamente inexistente: se trata de una variable continua.

Variable discreta. Es aquella en la que pueden distinguirse sin lugar a dudas 2 valores contiguos. **Variable continua.** Es aquella cuyos valores pueden medirse con tal precisión que la diferencia entre uno de sus valores y el siguiente puede perderse o ser insignificante.

Así, se puede hablar de variables aleatorias continuas y de variables aleatorias discretas, y de estas últimas es de las que trata este capítulo.

5.2 Distribuciones de probabilidad de variables aleatorias discretas

Como en las mediciones discretas se distinguen claramente los distintos valores de la variable, es posible construir una tabla para enlistar esos distintos valores que la variable puede asumir. Si se puede determinar la probabilidad de ocurrencia de cada uno de esos valores y se les incluye a ambos en una tabla, se tiene la **distribución de probabilidad para una variable aleatoria discreta**.

Distribución de probabilidad para una variable aleatoria discreta. Tabla que enlista todos los valores que una variable puede asumir con su probabilidad de ocurrencia correspondiente.

Por ejemplo, si se sabe, de acuerdo con los registros históricos, que la proporción de artículos defectuosos que fabrica una máquina que produce botellas es de 3%, se puede suponer que, si se extraen botellas al azar de las fabricadas por esta máquina, la probabilidad de que se trate de botellas defectuosas es de 3%. Al mismo tiempo se sabe que la probabilidad de que sea una botella sin defectos es de 97%. Con estos datos se puede construir una tabla de la distribución de probabilidad de esta variable aleatoria discreta:

Variable X	Probabilidad $P(X)$
Botella defectuosa	0.03
Botella sin defectos	0.97

Esta tabla, que muestra el conjunto de todos los resultados posibles (el espacio muestral, según se vio en el capítulo anterior), junto con sus correspondientes probabilidades, es una distribución de probabilidad y, como es una variable discreta, se trata entonces de una distribución de probabilidad de una variable discreta. También es **binomial** porque la variable de interés (botella defectuosa o no) sólo puede asumir 2 valores, se tiene, entonces, un ejemplo de una distribución de probabilidad de una variable binomial.

Variable binomial. Es aquella que sólo puede asumir 2 valores.

Se podría escoger al azar una sola botella para el examen (una única repetición del experimento) o se podrían escoger tantas botellas como fueran necesarias para el propósito del estudio. Este número de repeticiones del experimento se representa mediante la letra minúscula n. En el ejemplo anterior se supone que se considera una sola botella a la vez, por lo que $n = 1$. Si se escogieran 2 botellas, la distribución de probabilidad tendría la siguiente forma (en la sección sobre la distribución binomial se revisa la forma en la que se calculan estas probabilidades):

Variable Núm. de botellas defectuosas *X*	*P*(*X*)
0	0.9409
1	0.0582
2	0.0009

En donde 0, 1 y 2 representan todos los resultados posibles de número de botellas defectuosas en un lote de 2 de ellas.

Así, una distribución de probabilidad presenta el conjunto de las probabilidades asociadas a cada uno de todos los posibles valores de la variable aleatoria. En este tipo de estudios, a la probabilidad "de éxito" se le identifica mediante la letra p, que en este caso es de 0.03. La expresión "de éxito" no tiene el significado que comúnmente se le da, sino que se utiliza simplemente para identificar la probabilidad del suceso de interés. A la probabilidad del evento que no es el de interés sino el complementario se le conoce como *probabilidad "de fracaso"* y se le identifica mediante q. Así, en el ejemplo anterior, $p = 0.03$ y $q = 0.97$. Además, como la suma de las probabilidades de todos los sucesos posibles debe ser igual a 1, entonces, en un experimento binomial, la suma de p y q debe ser siempre igual a 1.

Aparte de las distribuciones de probabilidad binomial, existen otras distribuciones discretas de probabilidad, como la distribución multinomial, que implica una variable con más de 2 resultados posibles. Además de las 2 anteriores, en las secciones siguientes se revisan la distribución hipergeométrica y la de Poisson.

En la tabla 5.1 se resumen las características de las 4 distribuciones discretas de probabilidad que se analizan en este capítulo.

Tabla 5.1 Características de las principales distribuciones discretas de probabilidad

Distribución	Núm. de resultados posibles	Probabilidad
Binomial	2 resultados posibles	Igual *p* en cada ensayo
Poisson	2 resultados posibles	Igual *p* en cada ensayo
Multinomial	Más de 2 resultados	Igual *p* en cada ensayo
Hipergeométrica	2 resultados posibles	*p* diferentes en cada ensayo

Como puede verse en el cuadro anterior, las diferencias básicas entre estas 4 distribuciones se dan en términos del número de resultados posibles y de si la probabilidad de ocurrencia cambia o no en cada ensayo. Aunque se abundará en ello en las correspondientes subsecciones, conviene tener presente de entrada las principales diferencias en relación con la distribución binomial:

1. Entre la binomial y la multinomial la diferencia está en el número de resultados posibles.
2. Entre la binomial y la hipergeométrica la diferencia radica en que en la primera las probabilidades permanecen constantes, mientras que en la hipergeométrica las probabilidades cambian de un ensayo a otro; como se verá, esta distribución se relaciona estrechamente con el muestreo sin reemplazo, como cuando se extrae un naipe de una baraja y no se devuelve al mazo.
3. Entre la binomial y la de Poisson la diferencia es un tanto más sutil, ya que en ambas se tienen 2 resultados posibles e iguales probabilidades en todos los ensayos. En primer lugar, se verá más adelante que la binomial trata de X éxitos en n ensayos, en tanto que la Poisson aborda situaciones de promedios o valores esperados de éxitos en un continuo de tiempo o de espacio. Por ejemplo, la distribución binomial se aplica cuando se trata de determinar la probabilidad de obtener X número de artículos defectuosos en una muestra aleatoria de n de los producidos, y se conoce la probabilidad de que cualquiera de ellos sea defectuoso. Por su parte, la distribución de Poisson trata de problemas de, por ejemplo, el número de llamadas telefónicas recibidas por minuto en un conmutador, o del número de clientes que llegan por hora a la cola de la caja de un establecimiento comercial, o del número de defectos por metro de tela de un rollo.

Además, y como también se verá más adelante, la distribución de Poisson es el límite de la binomial cuando n es muy grande y p es tan pequeña que es precisamente cuando resulta más conveniente utilizar aquélla y no la binomial, pues es más fácil calcular la distribución de Poisson y es igualmente precisa.

En la sección siguiente se revisa la forma en la que se calculan la media y la desviación estándar de una distribución de probabilidades, incluyendo ejemplos de cómo se pueden utilizar las 3 funciones de Excel que calculan probabilidades para este tipo de distribuciones: Distr.Binom, Poisson y Distr.Hipergeom.

5.3 Media y varianza de una distribución de probabilidades

Se puede ilustrar el cálculo de la media y la varianza de una distribución de probabilidad mediante un ejemplo.

Ejemplo 5.1

En las 2 primeras columnas de la tabla siguiente se muestra el número de automóviles que se vendieron en una distribuidora de automóviles los últimos 500 días.

Núm. de autos vendidos por día X	Núm. de días (frecuencia absoluta) F	Frecuencia relativa	$P(X)$
0	90	0.18	0.18
1	200	0.40	0.40
2	100	0.20	0.20
3	80	0.16	0.16
4	25	0.05	0.05
5	5	0.01	0.01
Totales	500	1	1

Así, por ejemplo, el primer renglón señala que en cada uno de esos 90 días no se vendió ningún automóvil. La columna de frecuencia relativa se determinó de la misma manera en que se revisó antes: dividiendo la frecuencia absoluta entre el total de las frecuencias. Por ejemplo, el 0.18 del primer renglón se calculó dividiendo 90 entre 500, cociente que es, precisamente, 0.18 y es la frecuencia relativa correspondiente a la venta de 0 autos.

Como es importante aquí repasar el concepto de la frecuencia relativa, conviene recordar que, por ejemplo, el 0.40 del segundo renglón de esta columna significa que 40% de los días sólo se vendió un automóvil. Además, se repitió la columna de la frecuencia relativa con el encabezado de $P(X)$, es decir, la probabilidad de ocurrencia de cada valor de X porque, tal como se vio antes, se puede interpretar la frecuencia relativa como la probabilidad de ocurrencia del resultado correspondiente. En otras palabras, la misma interpretación que se hizo antes de la frecuencia relativa: "40% de los días sólo se vendió un automóvil" se puede interpretar también en términos de probabilidad diciendo que, si se elige un día al azar de entre los 500 días considerados, la probabilidad de que en ese día específico sólo se vendiera un automóvil es de 40 por ciento.

Resumiendo, en la siguiente tabla se muestra la distribución de probabilidad correspondiente a este ejemplo. A esta tabla se le añadieron las columnas necesarias para el cálculo de la media y de la desviación estándar, cuyos cálculos se detallan en los párrafos siguientes.

Núm. de autos vendidos por día X_i	$P(X_i)$	$X_i\,P(X_i)$	$X_i - \mu$	$(X_i - \mu)^2$	$(X_i - \mu)^2 \cdot P(X_i)$
0	0.18	0	–1.53	2.3409	0.421362
1	0.40	0.4	–0.53	0.2809	0.112360
2	0.20	0.4	0.47	0.2209	0.044180
3	0.16	0.48	1.47	2.1609	0.345744
4	0.05	0.2	2.47	6.1009	0.305045
5	0.01	0.05	3.47	12.0409	0.120409
Totales	1	1.53		23.1454	1.3491

Recordando la fórmula para calcular la media aritmética de una serie de datos y frecuencias:

$$\bar{X} = \frac{\sum X_i f_i}{\sum f_i} = \frac{\sum X_i f_i}{n}$$

Como en el caso de una distribución de probabilidad el conjunto de los resultados posibles constituye un universo (una población), en primer lugar se sustituye el símbolo del estadístico muestral $\bar{X}$, por el correspondiente al parámetro, la media de una población, μ, y, como se usa la frecuencia relativa como medida de la probabilidad, se sustituye f por $P(X)$, con lo cual la fórmula se convierte en:

$$\mu = \frac{\sum P(X_i) \cdot X_i}{\sum P(X_i)}$$

Recordando uno de los axiomas de la probabilidad, el que dice que la suma de las probabilidades de todos los resultados posibles mutuamente excluyentes y colectivamente exhaustivos de un experimento aleatorio es igual a 1, se tiene que:

$$\sum P(X_i) = 1$$

Entonces, la fórmula para el cálculo de la media de una distribución de probabilidad se convierte en:

$$\mu = \sum X_i \cdot P(X_i)$$

Además, vale la pena anotar que esta media de una distribución de probabilidad se le conoce como el "valor esperado de la variable", se le suele representar como $E(X)$. Resumiendo todo lo anterior en una sola ecuación, se tiene:

$$E(X) = \mu = \sum X_i \cdot P(X_i) \tag{5.1}$$

Volviendo al ejemplo, la media de la distribución de probabilidad es, de acuerdo con los cálculos que se anotan en la tabla anterior:

$$E(X) = \mu = \sum X_i \cdot P(X_i) = 1.53$$

Por su parte, la varianza de la distribución de probabilidad es:

$$Var(X) = \sigma^2 = \sum_{i=1}^{n} (X_i - \mu)^2 \cdot P(X_i) \tag{5.2}$$

la cual es equivalente a la fórmula para el cálculo de la varianza de una serie de datos y frecuencias que se vio en el capítulo 3, con los mismos cambios que se anotaron antes para la media aritmética: el denominador es también igual a 1, por lo que desaparece; se sustituyen las frecuencias por las probabilidades y se sustituye el símbolo del estadístico muestral s^2 por el correspondiente símbolo del parámetro, σ^2. Desprendiendo los cálculos de la tabla anterior, se tiene que

$$Var(X) = \sigma^2 = \sum_{i=1}^{n} (X_i - \mu)^2 \cdot P(X_i) = 1.3491$$

EJEMPLO 5.2

Se sabe que en intervalos de 15 minutos tomados aleatoriamente, el número de clientes que llegan a una cafetería sigue la distribución de probabilidades que se muestra en la siguiente tabla.

a) Calcule la media.
b) Calcule la varianza.
c) ¿Cuál es la probabilidad de que en un intervalo dado lleguen de 2 a 5 clientes?

Clientes X_i	$P(X_i)$
1	0.04
2	0.15
3	0.2
4	0.25
5	0.19
6	0.1
7	0.05
8	0.02
Total	1

Solución:

Clientes X_i	$P(X_i)$	$X_i \cdot P(X_i)$	$X_i - \mu$	$(X_i - \mu)^2$	$(X_i - \mu)^2 \cdot P(X_i)$
1	0.04	0.04	–3	9	0.36
2	0.15	0.3	–2	4	0.6
3	0.2	0.6	–1	1	0.2
4	0.25	1	0	0	0
5	0.19	0.95	1	1	0.19
6	0.1	0.6	2	4	0.4
7	0.05	0.35	3	9	0.45
8	0.02	0.16	4	16	0.32
Total	1	4			2.52

a) $E(X) = \mu = \sum X_i \cdot P(X_i) = 4$

b) $Var(X) = \sigma^2 = \sum_{i=1}^{n} (X_i - \mu)^2 \cdot P(X_i) = 2.52$

c) $P(2 \le X \le 5) = 0.15 + 0.2 + 0.25 + 0.19 = 0.79$

EJERCICIOS 5.3 Media y varianza de una distribución de probabilidades

1. Se muestra a continuación la distribución de los autos que llegan a un estacionamiento por hora.

Autos X_i	$P(X_i)$
0	0.05
1	0.1
2	0.15

Autos X_i	$P(X_i)$
3	0.25
4	0.3
5	0.1
6	0.05
Total	1

a) Calcule la media de la distribución.
b) Calcule la varianza.

2. Se sabe que la venta de una revista publicada quincenalmente sigue la distribución de probabilidad que se muestra en la siguiente tabla.

Revistas en miles X_i	$P(X_i)$
10	0.1
20	0.25
30	0.2
40	0.25
50	0.15
60	0.05
Total	1

a) Calcule la media.
b) Calcule la desviación estándar.
c) ¿Cuál es la probabilidad de que se vendan a lo mucho 30 000 revistas?

3. Una compañía de televisión de paga está por instalar un nuevo sistema de transmisión; en la siguiente tabla se muestra la posibilidad de que el servicio no esté disponible durante ciertos periodos por semana en la etapa inicial de instalación.

Periodos X_i	$P(X_i)$
2	0.01
3	0.08
4	0.3
5	0.41
6	0.13
7	0.07
Total	1

a) Calcule el número esperado de veces por semana en que el servicio no esté disponible.
b) Calcule la varianza.

4. Un agente de ventas realiza 10 vistas diarias y descubrió que la probabilidad de realizar cierto número de ventas está descrita por la siguiente distribución.

Ventas X_i	$P(X_i)$
1	0.04
2	0.15
3	0.25
4	0.31
5	0.2
6	0.05
Total	1

a) Calcule la media.
b) Calcule la varianza.

5. En la tabla siguiente se muestra el número de periódicos que vende un voceador en una importante avenida durante un periodo de 30 días.

Ventas	7	8	9	10	11	12	Suma
Días	2	4	8	7	5	4	30

a) Determine las probabilidades asociadas a cada evento.
b) Calcule la media de la distribución de probabilidades.
c) Calcule su varianza.

5.4 Distribución binomial

La **distribución binomial** es una distribución discreta de probabilidad que se basa en un experimento aleatorio que tiene las 3 características siguientes:

Distribución binomial. Distribución discreta de probabilidad que se basa en un experimento aleatorio.

1. La variable es binomial y, por lo tanto, discreta. Sólo hay 2 resultados posibles.
2. Las probabilidades de éxito o de fracaso no varían de una repetición a otra del experimento, o en otras palabras, los ensayos son independientes entre sí.
3. El experimento se lleva a cabo durante n ensayos iguales.

A un experimento como éste se le conoce también como proceso *Bernoulli* en honor del matemático suizo James Bernoulli (1654-1705), que hizo importantes contribuciones a la teoría de la probabilidad.

Cabe recordar que la suma de las probabilidades de todos los resultados posibles de un experimento aleatorio es 1 y, como en este caso sólo se tienen 2 resultados posibles, también su suma es igual a 1. Se

suele representar con p a la probabilidad de éxito o del resultado de interés y con q a su complemento y se puede representar esta relación como:

$$q = 1 - p \tag{5.3}$$

Se presenta en el siguiente ejemplo el razonamiento del que se deduce la fórmula para el cálculo de probabilidades binomiales.

EJEMPLO 5.3

En determinado juego de azar con 2 resultados posibles, la probabilidad de éxito es de 0.7, $p = 0.7$ y la probabilidad de fracaso es de 0.3, $q = 0.3$. Si se repite el juego 5 veces, ¿cuál es la probabilidad de tener éxito 3 veces?

En este caso, $n = 5$ y el número de casos de interés se representa como x y entonces $x = 3$. Se satisfacen las características de un experimento binomial porque hay sólo 2 resultados posibles y porque ningún resultado depende de los anteriores.

En primer lugar, es fácil ver que se pueden conseguir los 3 éxitos en secuencias diferentes: EEEFF, EEFEF, etc., y de acuerdo con la regla de multiplicación de probabilidades para eventos independientes:

$$P(EEEFF) = (0.7)(0.7)(0.7)(0.3)(0.3) = (0.7)^3\,(0.7)^2 = 0.343(0.09) = 0.03087$$
$$P(EEFEF) = (0.7)(0.7)(0.7)(0.3)(0.3) = (0.7)^3\,(0.7)^2 = 0.343(0.09) = 0.03087$$
$$P(EEFFE) = (0.7)(0.7)(0.7)(0.3)(0.3) = (0.7)^3\,(0.7)^2 = 0.343(0.09) = 0.03087$$

La probabilidad de tener 3 éxitos y, consecuentemente, 2 fracasos, es de 0.03087 y estos 3 éxitos se pueden obtener en diferente orden, pero siempre con la misma probabilidad. En símbolos:

$$P(EEEFF) = P(EEFEF) = P(EEFFE) = 0.03087$$

De aquí se desprende que esta probabilidad se calcula como:

$$P(EEEFF) = P(EEFEF) = P(EEFFE) = P(E)^3 \cdot P(E)^2$$

Si se sustituye $P(E)$ por p y $P(F)$ por q, se tiene que

$$P(EEEFF) = P(EEFEF) = P(EEFFE) = p^3q^2$$

Para encontrar la probabilidad de 3 éxitos y 2 fracasos, es necesario sumar las probabilidades de todas las formas en las que se puede obtener esa combinación de 3 y 2. El número de formas en las que sucederán esos 3 éxitos y 2 fracasos está dado por el número de combinaciones de 5 elementos tomados de 3 en 3, o en la simbología que se revisó en el tema de análisis combinatorio: C_3^5

$$C_3^5 = \frac{n!}{x!(n-x)!} = \frac{5 \cdot 4 \cdot 3 \cdot 2}{(3 \cdot 2)(2)} = \frac{5 \cdot 4 \cdot \cancel{3}}{\cancel{3} \cdot 2} = 10$$

Entonces,

$$P(3 \text{ éxitos y } 2 \text{ fracasos}) = 10(0.03087) = 0.3087 \text{ o } 30.87\%$$

Resumiendo, la distribución binomial tiene la siguiente función:

$$P(x) = C_n^n p^x q^{n-x} = \frac{n!}{x!(n-x)!} p^x q^{n-x} \tag{5.4}$$

EJEMPLO 5.4

Volviendo al proceso que se mencionó antes, al inicio de la sección 2, en el cual se tiene una máquina que, de acuerdo con registros históricos, produce 3% de artículos defectuosos, si se extraen al azar 2 artículos de los que se fabrican, las probabilidades de que haya 0 artículos defectuosos, 1 o 2, son:

$$P(0) = \binom{n}{x} p^x q^{n-x} = \binom{2}{0}(0.03)^0(0.97)^2 = 0.97^2 = 0.9409$$

$$(1) = \frac{n!}{x!(n-x)!} p^x q^{n-x} = \frac{2!}{1!1!}(0.03)^1(0.097)^1$$

$$= 2(0.03)(0.97) = 0.0582$$

$$P(2) = \binom{n}{x} p^x q^{n-x} = \binom{2}{2}(0.03)^2(0.97) = 0.03^2 = 0.0009$$

Como estos 3 eventos conforman la totalidad del espacio muestral de este experimento binomial, se puede verificar que la suma de sus probabilidades es igual a 1:

$$0.9409 + 0.0582 + 0.0009 = 1$$

Hay muchas situaciones que se pueden considerar ensayos binomiales:

- El lanzamiento de una moneda.
- Los artículos fabriles (televisiones, camisas, desodorantes, etc.), calificados como defectuosos o no defectuosos, si las probabilidades de que caigan en cada categoría no cambian y si el resultado en cada artículo no depende de los resultados de otros artículos.
- El resultado de una encuesta de opinión en la que se pide a las personas calificar artículos con criterios como “me gusta” o “no me gusta”.

Son importantes las características de los ensayos binomiales porque coinciden con el muestreo con reemplazo. Si se extraen bolas de una urna que contiene 6 pelotas rojas y 4 negras, y cada vez que se extrae una se devuelve a la urna después de anotar su color, las probabilidades de los 2 colores permanecen inalteradas en P(roja) = 0.6 y P(negra) = 0.4. Sin embargo, si se realiza el muestreo sin reemplazo, es decir, si no se devuelve la bola a la urna después de cada extracción, entonces se alteran las probabilidades, según sea el resultado que se obtiene.

Por ejemplo, si en la primera extracción sale una bola roja, y no se devuelve a la urna, para la segunda extracción habría 5 rojas y 4 negras, por lo que las nuevas probabilidades serían $P(\text{roja}) = \frac{5}{9} = 0.56$ y $P(\text{negra}) = \frac{4}{9} = 0.44$. Es evidente que, en este caso, sí se alteran las probabilidades en los sucesivos ensayos y, por lo tanto, no se trata de un ensayo binomial; la distribución de muestreo que se aplica en estos casos en los que no hay reemplazo es la distribución hipergeométrica, que se analiza más adelante.

Ejemplo 5.5

Un cerrajero se dio cuenta de que 1 de cada 8 de sus clientes regresan a reclamarle que sus llaves no funcionan. Si durante cierto día recibe a 7 clientes, ¿cuál es la probabilidad de que 3 le hagan una reclamación?

$$P\left(X = 3 \middle| n = 7, \quad p = 0.125\right) = C_3^7 (0.125)^3 (0.875)^4$$

$$= \frac{7!}{3!(4)!}(0.125)^3 (0.875)^4 = 0.04$$

Esta última notación incluye todos los datos en el planteamiento de la probabilidad:

$$P\left(X = 3 \middle| n = 7, \quad p = 0.125\right) = C_3^7 p^3 q^4$$

Muestra que se busca la probabilidad de 3 reclamos, dada una n de 7 clientes y una probabilidad de reclamo del 0.125, o 12.5 por ciento.

Ejemplo 5.6

En relación con el ejemplo anterior, ¿cuál es la probabilidad de que en un día reciba entre 4 y 6 quejas?

En este caso, es necesario calcular las probabilidades de 4, 5 y 6 quejas y sumarlas:

$$P\left(X = 4 \middle| n = 7, \quad p = 0.125\right) = C_4^7 (0.125)^4 (0.875)^3$$

$$= \frac{7 \cdot 6 \cdot 5 \cdot 4}{4 \cdot 3 \cdot 2}(0.125)^4 (0.875)^3 = 0.00572$$

$$P\left(X = 5 \middle| n = 7, \quad p = 0.125\right) = C_5^7 (0.125)^5 (0.875)^2$$

$$= \frac{7 \cdot 6 \cdot 5 \cdot 4 \cdot 3}{5 \cdot 4 \cdot 3 \cdot 2}(0.125)^5 (0.875)^2 = 0.00049$$

$$P\left(X = 6 \middle| n = 7, \quad p = 0.125\right) = C_6^7 (0.125)^6 (0.875)$$

$$= \frac{7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2}{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2}(0.125)^6 (0.875)^1 = 0.0000234$$

Finalmente:

$$(P(4 \cup 5 \cup 6) = P(4 \leq X \leq 6) = P(4) + P(5) + P(6)$$

$$= 0.00572 + 0.00049 + 0.0000234 = 0.0062334$$

Cabe recordar que la forma abreviada de resolver la fórmula de las combinaciones es extender el factorial de n hasta el x-avo término y dividir este producto entre el factorial de x.

5.4.1 Media y varianza de la distribución binomial

Tal y como se vio en la sección 5.3 es posible calcular la media aritmética y la desviación estándar de cualquier distribución de probabilidad siguiendo los procedimientos que se revisaron repetidamente y que las distribuciones de probabilidad se pueden resumir mediante las siguientes fórmulas cuya media aritmética es:

$$E(X) = \mu = \Sigma X_i \cdot P(X_i) \tag{5.1}$$

$$Var(X) = \sigma^2 = \sum_{i=1}^{n}(X_i - \mu)^2 \cdot P(X_i) \tag{5.2}$$

Sin embargo, y por otra parte, se derivaron expresiones más sencillas para determinar estos 2 parámetros para la distribución binomial, y son:

Media aritmética:

$$\mu = np \tag{5.5}$$

Su varianza es:

$$\sigma^2 = npq \tag{5.6}$$

Se muestra en el ejemplo siguiente que cualquiera de los 2 procedimientos conduce a los mismos resultados.

EJEMPLO 5.7

La probabilidad de que un edificio seleccionado aleatoriamente esté asegurado es de 0.38. Calcular el espacio muestral de las probabilidades de que los edificios de una muestra de 3 edificios estén asegurados.

Solución: Recordando que el espacio muestral es el conjunto de todos los resultados posibles, mutuamente excluyentes y colectivamente exhaustivos de un experimento aleatorio, en este ejemplo, los resultados posibles de la variable aleatoria "número de edificios asegurados" en una muestra de 3 de ellos es: 0, 1, 2 y 3. En la tabla siguiente se muestran las probabilidades binomiales asociadas a cada resultado:

X_i	$P(X_i)$	$X_i \cdot P(X_i)$	$X_i - \mu$	$(X_i - \mu)^2$	$(X_i - \mu)^2 \cdot P(X_i)$
0	0.24	0.00	–1.14	1.30	0.3097
1	0.44	0.44	–0.14	0.02	0.0086
2	0.27	0.54	0.86	0.74	0.1986
3	0.05	0.16	1.86	3.46	0.1898
	1	1.14			0.7068

Las probabilidades de cada posible valor X se calcularon de acuerdo con la probabilidad binomial y las columnas restantes son parte de los cálculos que ya se repasaron antes para determinar la media y la desviación estándar, de manera que:

$$\mu = 1.14, \text{ y } \sigma = \sqrt{0.7068} = 0.8407$$

De la misma manera:

$$\mu = np = 3(0.38) = 1.14, \text{ y}$$

$$\sigma^2 = npq = 3(0.38)(0.62) = 0.7068$$

$$\sigma = \sqrt{0.7068} = 0.8407$$

Estos resultados son, precisamente, lo que se deseaba ilustrar. Y, aunque no es materia de este texto abundar sobre este tema, se puede apreciar fácilmente que conviene conocer estas características de la distribución binomial ($\mu = np$, y $\sigma = npq$) porque, entonces, para calcular su media y su varianza no es necesario seguir el procedimiento de cálculo completo, sino que se pueden utilizar estas fórmulas que abrevian considerablemente las operaciones.

5.4.2 Distribución binomial, tablas de probabilidades binomiales y Excel

Las tablas de probabilidades binomiales son especialmente útiles cuando se requiere calcular muchas probabilidades o, también, para verificar que la suma de las probabilidades de un espacio muestral binomial suman 1, con lo que se puede verificar que no haya errores en los cálculos.

Sin embargo, por otra parte, con las potentes calculadoras con las que se cuenta en la actualidad, resolver la función binomial es muy sencillo, y además Excel cuenta con una función para hacer estos cálculos que es muy fácil de usar. Las aplicaciones de la distribución binomial se pueden resolver fácilmente con esta función de Excel y tiene la siguiente sintaxis:

DISTR.BINOM(Número de éxitos,número de ensayos,probabilidad,acumulado)

Los primeros 3 parámetros se explican por sí solos. El último, "acumulado", se usa para determinar el resultado del cálculo: si no se anota nada (aunque sí se debe anotar una coma después del valor de la probabilidad) o se anota "FALSO", entonces la función devuelve el valor de la probabilidad para los parámetros anotados; si se anota "VERDADERO" entonces el valor que se produce es la probabilidad acumulada desde cero y hasta el valor anotado de X, el número de éxitos.

En el ejemplo 5.3, que trata de una máquina que produce 3% de artículos defectuosos, se calcularon las probabilidades de que haya 0, 1 o 2 defectuosos en una muestra de 2 de ellos. Con Excel:

=DISTR.BINOM(0,2,.3.0,)
=DISTR.BINOM(1,2,.3.0,)
=DISTR.BINOM(2,2,.3.0,)

se obtienen los mismos resultados, 0.9409, 0.0582 y 0.0009, que se calcularon antes.

En el ejemplo 5.4, que trata del cerrajero en el que 1 de cada 8 clientes regresa con una reclamación, se pregunta cuál es la probabilidad de que de 7 clientes, 3 le reclamen y se encontró que es de 0.04. Con la función de Excel =DISTR.BINOM(3,7,.125.0,) se obtiene el mismo resultado.

En el ejemplo 5.5 se pide, utilizando los mismos datos del cerrajero, calcular la probabilidad de que tenga entre 4 y 6 quejas. Se puede resolver esta pregunta con Excel simplemente al calcular las probabilidades de 4, 5 y 6 quejas y sumarlas. Para ilustrar cómo se pueden calcular todas las probabilidades de manera sencilla, se pueden anotar los valores 0, 1, … , 7 en los primeros 8 renglones de la columna A de una hoja de Excel y luego anotar la función =DISTR.BINOM(A1,7,.125.0,) en la celda B1 para obtener la probabilidad correspondiente a 0 reclamaciones y luego simplemente copiar esa misma fórmula hasta el renglón A7 para obtener todo el espacio muestral, que es:

X	*P*(*X*)
0	0.392695904
1	0.392695904
2	0.168298244
3	0.040071011
4	0.005724430
5	0.000490665
6	0.000023365
7	0.000000477
	1

De esta distribución de probabilidad se pueden extraer las probabilidades de 4, 5 y 6 quejas y sumarlas para llegar al mismo resultado anterior, salvo diferencias menores por redondeo.

La tabla 1 del apéndice es una tabla de probabilidades binomiales para valores seleccionados de sus parámetros, construida con esta función Distr.Binom de Excel. Se le incluye aquí porque puede resultar útil pero es importante resaltar que la cantidad de valores diferentes para los parámetros que se requieren para calcular las probabilidades (n, x y p) son limitados en la tabla por cuestiones de espacio, en tanto que con una calculadora se puede determinar la probabilidad para cualquier valor que se requiera y con más facilidad con la función de Excel.

ejercicios 5.4 Distribución binomial

1. La probabilidad de que un cliente potencial elegido al azar realice una compra es de 0.20. Si un agente de ventas visita 6 clientes, ¿cuál es la probabilidad de que realice exactamente 4 ventas?

2. En una empresa, la probabilidad de que un empleado participe en el programa de caja de ahorro es de 65%. Si se eligen 5 empleados al azar, ¿cuál es la probabilidad de que al menos 2 de ellos participen en el programa?

3. Una casa de empeño informó que 30% de los préstamos garantizados con joyería vencieron. Si se toma una muestra aleatoria de 4 préstamos, ¿cuál es la probabilidad de que:

a) ninguno esté vencido?
b) exactamente 2 estén vencidos?
c) de 2 a 4 estén vencidos?

4. El 0.35 de los trabajadores en una planta están conformes con la dirección. Se toma una muestra de 10 personas a las que se les realiza una encuesta anónima. ¿Cuál es la probabilidad de que la mitad de los interrogados estén conformes con la dirección de la planta?

5. Una compañía comercializadora de alimento para perro creó una nueva estrategia de promoción de ventas que tiene 10% de posibilidades de ser exitosa. Si se aplica en 20 tiendas, ¿cuál es la probabilidad de que:

a) no tenga éxito en ninguna de las tiendas?
b) hasta en 3 tiendas tenga éxito?
c) en las 20 tiendas tenga éxito?

Media y varianza de una distribución de probabilidad binomial

6. La probabilidad de encontrar un pantalón con algún defecto de la producción total diaria de una maquiladora es de 0.24.

a) Calcule el espacio muestral de probabilidades de que los pantalones estén defectuosos de una muestra de 4.
b) Obtenga la media de la distribución de probabilidad.
c) Calcule su varianza por los 2 métodos, a partir de la distribución de probabilidad y de la fórmula para la varianza.

7. La probabilidad de que un niño repruebe 1o. de primaria es de 0.1.

a) Calcule el espacio muestral de probabilidades de que se repruebe 1o. de primaria de una muestra de 5 niños.
b) Obtenga la media de la distribución de probabilidad.
c) Calcule su varianza por los 2 métodos a partir de la distribución de probabilidad y de la fórmula para la varianza.

8. En electrónica, la probabilidad de que una patente seleccionada aleatoriamente tenga éxito en el mercado es de 0.08.

a) Calcule el espacio muestral de probabilidades de que las patentes de una muestra de 5 tengan éxito en el mercado.
b) Obtenga la media de la distribución de probabilidad.
c) Calcule su varianza por los 2 métodos, a partir de la distribución de probabilidad y de la fórmula para la varianza.

9. La probabilidad de que un detective, seleccionado aleatoriamente, resuelva un crimen en el primer mes de investigaciones es de 0.19. Calcule el espacio muestral de las probabilidades de que un detective resuelva en un mes los 4 casos de una muestra de crímenes.

10. En una secundaria la probabilidad de que un maestro, seleccionado aleatoriamente, llegue 15 minutos tarde a su clase es de 0.25. Calcule el espacio muestral de las probabilidades de que los maestros de una muestra de 6 lleguen hoy tarde a su clase.

5.5 Tres formas de presentar una distribución de probabilidad

Ahora que ya se revisó la distribución binomial se puede ilustrar este tema que aborda las 3 formas en que se puede presentar una distribución de probabilidad teórica:

1. Tabla.
2. Gráfica.
3. Función.

Volviendo al ejemplo anterior, el 5.6, en el que la probabilidad de que un edificio seleccionado aleatoriamente esté asegurado es de 0.38, y en donde se calculó el espacio muestral de las probabilidades de que los edificios de una muestra de 3 edificios tengan seguro. Se reproduce en seguida la tabla de la distribución de probabilidad para $n = 3$ y $p = 0.38$:

X_i	$P(X_i)$
0	0.24
1	0.44
2	0.27
3	0.05
	1

En esta tabla se aprecia que se puede presentar la distribución de probabilidad enumerando todos los resultados posibles (el espacio muestral del experimento) junto con sus correspondientes probabilidades.

En segundo lugar, estos mismos datos se pueden presentar en forma de gráfica, como la que se muestra en la figura 5.1.

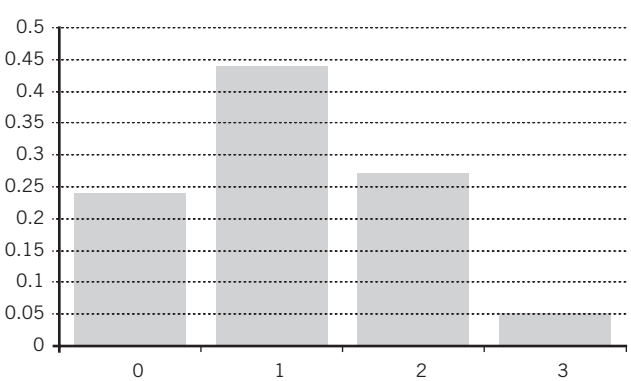


Figura 5.1 Gráfica de la distribución de probabilidad binomial para $n = 3$ y $p = 0.38$.

En tercer y último lugar, estos valores del espacio muestral y sus probabilidades se pueden presentar en términos de la función de probabilidad binomial:

$$P(0|n = 3, \quad p = 0.38) = \frac{3!}{0!(3-0)!}(0.38)^0(0.62)^3 = 0.24$$

$$P(1|n = 3, \quad p = 0.38) = \frac{3!}{1!(3-1)!}(0.38)^1(0.62)^2 = 0.4382$$

$$P(2|n = 3, \quad p = 0.38) = \frac{3!}{2!(3-2)!}(0.38)^2(0.62)^1 = 0.2686$$

$$P(3|n = 3, \quad p = 0.38) = \frac{3!}{3!(3-3)!}(0.38)^3(0.62)^0 = 0.0549$$

Ejemplo 5.8

La probabilidad de que en un concierto masivo de rock haya lesionados es de 12%. Calcule el espacio muestral de que haya lesionados en una muestra de 4 conciertos y presentarlos como:

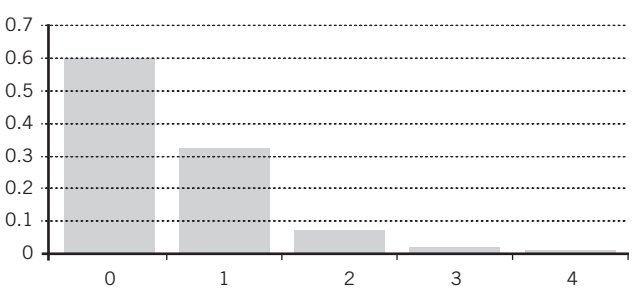
a) Tabla.
b) Gráfica de barras.
c) Función.

Solución:

a) Tabla.

X_i	$P(X_i)$
0	0.5999
1	0.3271
2	0.0669
3	0.0060
4	0.0002
	1.00

b) Gráfica de barras.



c) Función.

$$P(0|n = 4, \quad p = 0.12) = \frac{4!}{0!(4-0)!}(0.12)^0(0.88)^4 = 0.5999$$

$$P(1|n = 4, \quad p = 0.12) = \frac{4!}{1!(4-1)!}(0.12)^1(0.88)^3 = 0.3271$$

$$P(2|n = 4, \quad p = 0.12) = \frac{4!}{2!(4-2)!}(0.12)^2(0.88)^2 = 0.0669$$

$$P(3|n = 4, \quad p = 0.12) = \frac{4!}{3!(4-3)!}(0.12)^3(0.88)^1 = 0.0060$$

$$P(4|n = 4, \quad p = 0.12) = \frac{4!}{4!(4-4)!}(0.12)^4(0.88)^0 = 0.0002$$

Ejercicios 5.5 Tres formas de presentar una distribución de probabilidad

1. Unos delincuentes determinaron que, en los últimos 10 años, 13% de los robos a bancos tuvieron éxito. Si traman realizar 5 robos, calcule el espacio muestral de que tengan éxito y preséntelos como:

a) Tabla.
b) Gráfica de barras.
c) Función.

2. La probabilidad de que cierta estrella de béisbol conecte un *home run* es de 17%. Calcule el espacio muestral de que se conecten *home runs* en una muestra de 6 turnos al bate y preséntelo en forma de:

a) Tabla.
b) Gráfica de barras.
c) Función.

3. La probabilidad de que una familia de campesinos centroamericanos viaje a Estados Unidos y haga fortuna es de 19%. Calcule el espacio muestral en una muestra de 4 familias que buscan fortuna en las nuevas tierras y preséntelos en forma de:

a) Tabla.
b) Gráfica de barras.
c) Función.

4. La probabilidad de que doña Tere, una devota anciana de 78 años, falte a misa en domingo es de 8%. Calcule el espacio muestral de que falte a misa en una muestra de 7 domingos y preséntela en forma de:

a) Tabla.
b) Gráfica de barras.
c) Función.

5.6 Distribución de Poisson

La distribución de Poisson, otra distribución discreta, lleva el nombre del matemático francés Simeon Denis Poisson (1781-1840), quien publicó su derivación en 1837. Esta distribución está dada por la siguiente función:

$$P(x) = \frac{e^{-\lambda}\lambda^{x}}{x!} \tag{5.7}$$

Esta fórmula representa la probabilidad de que una variable aleatoria discreta X asuma el valor específico x, y e es la constante natural, que es igual a 2.71828, aproximadamente; λ, la letra griega lambda, representa la media de la distribución y está dada por $\lambda = np$. Adicionalmente, esta media de la distribución es igual a su varianza. Se ilustra esto más adelante, en la subsección 5.6.3.

La distribución de Poisson es útil para determinar probabilidades de fenómenos que ocurren en un continuo de espacio o de tiempo, como defectos en un rollo de tela, erratas en las páginas de un libro, número de llamadas telefónicas que llegan a un conmutador por intervalo o personas que llegan a hacer fila ante una ventanilla de atención al público. De hecho, este último caso representa una situación especial que se estudia con detenimiento en un área matemática que se denomina *investigación de operaciones*, en un tema que se conoce como "teoría de colas" o "teoría de líneas de espera", en el que el uso de la distribución de Poisson ocupa un lugar preponderante.

Las características de una variable aleatoria Poisson son:

- El experimento aleatorio consiste en contar el número de veces que ocurre el evento en una unidad determinada de espacio o de tiempo.
- Las ocurrencias de los eventos son mutuamente independientes.
- La probabilidad de ocurrencia es igual para todos los eventos.
- En una unidad de espacio o de tiempo muy reducida, la probabilidad de ocurrencia de más de un evento es tan pequeña que es prácticamente despreciable.

Se revisan en seguida algunos ejemplos.

EJEMPLO 5.9

Un libro de 500 páginas tiene 200 errores de impresión distribuidos aleatoriamente. Calcule la probabilidad de que cualquier página elegida al azar tenga un error.

Solución: Aquí, λ, la media de la distribución, es de:

$$\lambda = \frac{200}{500} = 0.40$$

O sea, que el libro tiene un promedio de 0.40 errores por página. Entonces,

$$P(1) = \frac{e^{-\lambda}\lambda^{x}}{x!} = \frac{2.71828^{-0.40}(0.40)^{1}}{1} = 0.2681$$

También, la probabilidad de que una página cualquiera tenga 2 errores es de:

$$P(2) = \frac{e^{-\lambda}\lambda^{x}}{x!} = \frac{2.71828^{-0.40}(0.40)^{2}}{2} = 0.0536$$

También podría ser de interés determinar cuál es la probabilidad de que cualquier página tenga 2 errores o menos (es decir, que tenga 2, 1 o 0 errores); para determinar esta probabilidad se requiere sumar las probabilidades de cada uno de esos 3 casos. Ya se calcularon antes $P(1)$ y $P(2)$. Ahora:

$$P(0) = \frac{e^{-\lambda}\lambda^{x}}{x!} = \frac{2.71828^{-0.40}(0.40)^{0}}{0!} = \frac{2.71828^{-0.40}(1)}{1} = 0.6703$$

Entonces, la probabilidad de que cualquier página tenga 2 o menos errores es de:

$$P(0) + P(1) + P(2) = 0.2681 + 0.0536 + 0.6703 = 0.992$$

Asimismo, se puede calcular, a partir de este último resultado, la probabilidad de que una página cualquiera tenga 3 errores o más como:

$$P(x \geq 3) = 1 - 0.992 = 0.008$$

5.6.1 Distribución de Poisson, tablas de probabilidades Poisson y Excel

La función de Excel que calcula probabilidades Poisson tiene la siguiente sintaxis:

=POISSON(*X*,media,acumulado)

en donde *X* es el número de sucesos, "media" es la media de la distribución y, al igual que con la función Distr.Binom, el último parámetro "acumulado" se usa para determinar el resultado del cálculo: si no se anota nada (aunque sí se debe anotar una coma después del valor de la probabilidad) o se anota "FALSO", entonces la función devuelve el valor de la probabilidad para los parámetros anotados; si se anota "VERDADERO", o "1", entonces el valor que se produce es la probabilidad acumulada desde 0 y hasta el valor anotado de *X*, el número de éxitos.

En el ejemplo anterior, el 5.9, la probabilidad de un error se calcula con Excel mediante =POISSON(1,4.0,) produciendo el mismo resultado de 0.2681.

Para encontrar todos los resultados que se calcularon antes, se podrían anotar los números 0 al 5 en los primeros 6 renglones de una hoja de Excel y, después, anotar =POISSON(A1,4.0,) en la celda B1 y luego copiar la fórmula hasta el renglón 6, con lo que se obtiene la siguiente distribución de probabilidad:

X	*P*(*X*)
0	0.670320
1	0.268128
2	0.053626
3	0.007150
4	0.000715
5	0.000057
6	0.000004
	1

Al igual que con las tablas de la distribución binomial, las tablas de probabilidades Poisson son especialmente útiles cuando se requiere calcular muchas probabilidades o, también, para verificar que la suma de las probabilidades de un espacio muestral tipo Poisson suman 1. La tabla 2 del apéndice es una tabla de probabilidades Poisson para valores seleccionados de sus parámetros, construida con esta función de Excel, Poisson.

Por otra parte, se incluye aquí la tabla porque puede resultar útil, pero es importante recordar que la cantidad de valores diferentes para los parámetros que se requieren para calcular las probabilidades (*X* y la media) son limitados en la tabla por cuestiones de espacio, en tanto que con una calculadora se puede determinar la probabilidad para cualquier valor que se requiera y, con mayor facilidad, con la función de Excel. A continuación, más ejemplos de esta distribución Poisson.

Ejemplo 5.10

En un estudio sobre accidentes laborales, la compañía Seguros Nacionales analiza los historiales de 20 años de 100 empresas, lo cual conduce a los siguientes 2 000 resultados de accidentes por año y por empresa:

X_i núm. de accidentes por año y por empresa	f_i frecuencia observada
0	1 100
1	600
2	270
3	20
4	10

Si se considera que los accidentes tienen una distribución aleatoria, se puede calcular con estos datos una media teórica y, a partir de ella, la distribución teórica del número de accidentes lo cual, a su vez, se puede utilizar en los cálculos actuariales para determinar el costo de pólizas de seguro para cubrir este tipo de accidentes.

La media de la distribución es:

X_i núm. de accidentes por año y por empresa	f_i frecuencia observada	$f(X_i)$
0	1 100	0
1	600	600
2	270	540
3	20	60
4	10	40
	Suma	1 240

Así, λ, la media de la distribución observada, es:

$$\lambda = \frac{1\,240}{2\,000} = 0.62 \text{ accidentes por empresa, por año.}$$

A partir de esta media teórica, la distribución teórica de la probabilidad de accidentes se calcula como:

$$P(0) = \frac{2.71828^{-0.62}(0.62)^0}{0!} = 0.53794$$

$$P(1) = \frac{2.71828^{-0.62}(0.62)^1}{1!} = 0.53794(0.62) = 0.33353$$

$$P(2) = \frac{2.71828^{-0.62}(0.62)^2}{2!} = \frac{0.53794(0.62)^2}{2} = 0.05562$$

$$P(3) = \frac{2.71828^{-0.62}(0.62)^3}{3!} = \frac{0.53794(0.62)^3}{3\cdot 2}$$

$$= \frac{0.53794(0.238328)}{6} = 0.02137$$

Y, finalmente,

$$P(4) = \frac{2.71828^{-0.62}(0.62)^4}{4!} = \frac{0.53794(0.62)^4}{4\cdot 3\cdot 2}$$

$$= \frac{0.53794(0.14776)}{24} = 0.00331$$

Resumiendo:

X_i núm. de accidentes por año y por empresa	f_i frecuencia observada	$P(X_i)$	Frecuencia relativa
0	1 100	0.53794	0.55
1	600	0.33353	0.30
2	270	0.05562	0.14
3	20	0.02137	0.01
4	10	0.00331	0.01
	Suma	0.95177	

Esta suma de 0.95177, o 95.18% significa que, teóricamente, la probabilidad de que en una empresa se den 5 o más accidentes en un año determinado es de:

$$P(x \geq 5) = 1 - 0.95177 = 0.04823$$

A partir de los datos de la tabla se puede saber, por ejemplo, que la probabilidad de que haya menos de 2 accidentes por año en una empresa determinada es de:

$$P(0) + P(1) = 0.53794 + 0.33353 = 0.87147, \text{ u } 87.15\%$$

Adicionalmente, si se observa la última columna de la tabla anterior, la de la frecuencia relativa, se puede apreciar que las frecuencias relativas son muy similares a las frecuencias teóricas obtenidas a partir de la distribución de Poisson, por lo que se confirma que la distribución del número de accidentes se aproxima a la de Poisson. En el capítulo sobre la distribución χ^2 (ji cuadrada), en la sección 11.9.2 del capítulo 11 se ve un tema sobre bondad de ajuste en el que se estudia la forma de evaluar si esta aproximación permite afirmar que, efectivamente, la distribución del número de accidentes sigue la distribución de Poisson.

Ejemplo 5.11

Una empresa de aviación determinó que 2% de las personas que reservan cierto vuelo no se presentan. Por ello, decide vender 149 lugares para un avión que tiene 148 asientos. Calcule la probabilidad de que todas las personas que se presenten tengan asiento disponible, suponiendo que las llegadas de los pasajeros siguen una distribución de Poisson.

Solución: En este caso,

$$\lambda = np = 150(0.02) = 3$$

sería el promedio de personas que no se presentan para un conjunto de reservaciones de 149. Las probabilidades de que no falte a la cita ninguna de las 149 personas, o que falte una son:

$$P(0) = \frac{e^{-\lambda}\lambda^x}{x!} = \frac{2.71828^{-3}(3)^0}{0!} = 0.04979$$

$$P(1) = \frac{e^{-\lambda}\lambda^x}{x!} = \frac{2.71828^{-3}(3)^1}{1!} = 0.14936$$

Como los únicos casos en los que no todos los pasajeros obtienen asiento es cuando no falta ninguno o sólo falta uno, la probabilidad de que todas las personas que se presenten tengan asiento disponible es:

$$1 - P(0) - P(1) = 1 - 0.04979 - 0.14936 = 0.8009, \text{ u } 80.09\%$$

Ejemplo 5.12

Durante una hora, una impresora saca en promedio 3 hojas por minuto, ¿cuál es la probabilidad de que la impresora imprima 5 hojas durante un minuto escogido al azar?

Solución: Como en este caso ya se tiene el dato de la media de la distribución, se procede a sustituir directamente en la fórmula:

$$P(5) = \frac{2.71828^{-3}(3)^5}{5!} = \frac{(0.04979)(243)}{12}$$

$$= \frac{12.09828}{120} = 0.1008$$

5.6.2 Distribución de Poisson como aproximación de la distribución binomial

En casos en los que aplica la distribución binomial y n es relativamente grande y p relativamente pequeña, se acostumbraba argumentar que los cálculos de la probabilidad binomial eran un tanto complicados y, por ello, se sugería utilizar la distribución de Poisson para aproximar los resultados porque, tal como puede verse, la fórmula es más sencilla de resolver. Sin embargo, con las herramientas de cálculo con que se disponen en la actualidad, desde calculadoras más o menos sencillas hasta computadoras, el cálculo de probabilidades binomiales es sencillo, independientemente de los valores de n y de p. Para ilustrar esta aproximación entre las 2 distribuciones se presenta en seguida un ejemplo.

Ejemplo 5.13

Los registros de una empresa muestran que la probabilidad de que falle cierto tipo de foco en las primeras 100 horas de uso es $p = 0.0005$. Determine la probabilidad de que, de un lote de 1 000 focos, cuando mucho 2 fallen en las primeras 100 horas de uso *a*) usando la distribución de Poisson y *b*) usando la binomial.

Solución: El "cuando mucho 2" implica que se busca la probabilidad de que fallen 0, 1 o 2 focos. Así, la respuesta es dada por $P(X = 0) + P(X = 1) + P(X = 2)$ y, como $n = 1\,000$ y $p = 0.0005$,

$$\mu = np = 1\,000(0.0005) = 0.5, \text{ y}$$

$$P(0) = \frac{2.71828^{-0.5}(0.5)^0}{0!} = \frac{0.6065(1)}{1} = 0.6065$$

$$P(1) = \frac{2.71828^{-0.5}(0.5)^1}{1!} = \frac{0.6065(0.5)}{1} = 0.30325$$

$$P(2) = \frac{2.71828^{-0.5}(0.5)^2}{2!} = \frac{0.6065(0.25)}{2} = 0.07581$$

Así, la probabilidad de que fallen cuando mucho 2 focos, siguiendo la distribución de Poisson, es:

$$P(X \le 2) = P(X = 0) + P(X = 1) + P(X = 2) = 0.6065 + 0.30325 + 0.07581 = 0.98556$$

Haciendo ahora los cálculos con la distribución binomial se tiene:

$$P(x) = \binom{n}{x} p^x q^{n-x} = \frac{n!}{x!(n-x)!} p^x q^{n-x}$$

$$P(0) = \frac{n!}{x!(n-X)!} p^x q^{n-x} = \frac{1\,000!}{0!1\,000!}(0.0005^0)(0.9995^{1\,000})$$

$$P(0) = (1)(0.9995^{1\,000}) = 0.6065$$

$$P(1) = \frac{n!}{x!(n-X)!}\left(p^X q^{n-X}\right) = \frac{1\,000!}{1!\,(1\,000-1)!}(0.0005^1)(0.9995^{999})$$

$$= (1\,000)(0.0005^1)(0.9995^{999}) = 0.3034$$

$$P(2) = \frac{n!}{x!(n - X)!} p^X q^{n-X} = \frac{1\,000!}{2!998!} 0.0005^2 \cdot 0.9995^{998}$$

$$= \frac{1\,000(999)}{2}(0.00000025)(0.9995^{998}) = 0.07581$$

$$P(x \leq 2) = P(x = 0) + P(x = 1) + P(x = 2)$$

$$= 0.6065 + 0.3034 + 0.07581 = 0.98571$$

Que son prácticamente los mismos resultados que se obtuvieron antes con la distribución de Poisson.

5.6.3 Media y varianza de la distribución de Poisson

Tal como se menciona al principio de esta sección 5.6, la media de la distribución de Poisson es

$\mu = E(X) = \lambda = np$, y es igual, también, a su varianza:

$$Var(X) = \lambda = np$$

EJEMPLO 5.14

En un banco se reciben 3 cheques sin fondos cada día. Si se elabora la distribución de probabilidad de Poisson con este dato se obtienen los resultados que se resumen en la tabla 5.2.

Tabla 5.2

X	$P(X)$	$X \cdot P(X)$	$(X - \bar{X})^2$	$(X - \bar{X})^2 \cdot P(X)$
0	0.0498	0.0000	9	0.4481
1	0.1494	0.1494	4	0.5974
2	0.2240	0.4481	1	0.2240
3	0.2240	0.6721	0	0.0000
4	0.1680	0.6721	1	0.1680
5	0.1008	0.5041	4	0.4033
6	0.0504	0.3025	9	0.4537
7	0.0216	0.1512	16	0.3457
8	0.0081	0.0648	25	0.2025
9	0.0027	0.0243	36	0.0972
10	0.0008	0.0081	49	0.0397
11	0.0002	0.0024	64	0.0141
12	0.0001	0.0007	81	0.0045
Sumas	1.0000	2.9998		2.9983

En la suma de la tercera columna se puede ver que la media de la distribución es prácticamente 3 (2.9998) y en donde la diferencia se debe a redondeo y a que no se consideraron todos los valores posibles de X, sino que se tomó hasta $X = 12$, el cual ya es muy cercano a 0. Además, en la suma de la quinta y última columnas se puede apreciar que la varianza es también prácticamente igual a 3. Con esto se ilustra cómo, para la distribución de Poisson:

$$E(X) = \mu = Var(X) = np \tag{5.8}$$

EJERCICIOS 5.6 Distribución de Poisson

1. En un centro telefónico de atención a clientes se reciben en promedio 5 llamadas por hora. ¿Cuál es la probabilidad de que en una hora seleccionada aleatoriamente se reciban exactamente 3 llamadas?
2. En una tienda departamental, en la sección de electrodomésticos, un promedio de 12 personas por hora le hacen preguntas al encargado. ¿Cuál es la probabilidad de que más de 3 personas se acerquen al encargado a hacer preguntas en un periodo de 10 minutos?
3. De las refacciones enviadas por un proveedor, 20% son defectuosas. Si se selecciona una muestra de 25 artículos, ¿cuál es la probabilidad de que más de 2 resulten defectuosas?
4. En promedio, 5 personas por hora realizan algún trámite en una oficina gubernamental. ¿Cuál es la probabilidad de que 5 o más realicen algún trámite en una hora en particular?
5. Cada rollo de 500 metros de tela tiene 2 defectos en promedio (rasguños, hilos sueltos). ¿Cuál es la probabilidad de que en un segmento de 100 metros en particular no exista defecto alguno?

Aproximación de probabilidades binomiales con la distribución de Poisson

6. Los analistas pronostican que 0.08% de las microempresas que se abran este año lograrán crecer a empresas medianas en 5 años. Determine la probabilidad de que en una muestra de 500 nuevas microempresas, solamente 3 logren crecer a medianas empresas,

   a) usando la distribución de Poisson y
   b) usando la binomial.

7. La probabilidad de que un maestro de Wushu tenga un error al ejecutar una forma de Taiji Quan en una exhibición es del 0.003. Si el maestro ejecutó 190 exhibiciones en los últimos 3 años, ¿qué probabilidad hay de que tuviera algún error en 4 exhibiciones? Resuelva:

   a) usando la distribución de Poisson y
   b) usando la binomial.

8. Un fabricante de hornos de microondas estima que la probabilidad de que un consumidor reclame la garantía por algún desperfecto del equipo durante el primer mes es de 0.005. Si este mes se vendieron 125 hornos, ¿cuál es la probabilidad de que se descompongan de 3 a 5 hornos? Resuelva:

   a) usando la distribución de Poisson y
   b) usando la binomial.

9. La probabilidad de que a un ciclista en una excursión por la montaña se le descomponga su bicicleta es de 0.009. Si este fin de semana fueron a la montaña 87 ciclistas, ¿cuál es la probabilidad de que se descompongan exactamente 7 bicicletas? Resuelva:

   a) usando la distribución de Poisson y
   b) usando la binomial.

10. Un laboratorio estima que la probabilidad de encontrar en un pozo de agua metales pesados en cantidades superiores a las permitidas por la ley, es de 0.007. Si analiza 200 pozos, ¿cuál es la probabilidad de que de 9 a 11 pozos contengan agua con más metal que lo permitido? Resuelva:

   a) usando la distribución de Poisson y
   b) usando la binomial.

Media y desviación estándar de la distribución de Poisson

11. En un restaurante hay al día, en promedio, 5 reservaciones a comer. Elabore la distribución de probabilidad de Poisson con este dato y verifique si se cumple que $E(X) = Var(X) = np$.
12. En un gran hospital, cada hora en promedio operan a 4 personas. Elabore la distribución de probabilidad de Poisson con este dato y verifique si se cumple que $E(X) = Var(X) = np$.
13. Un arqueólogo espera encontrar 8 huesos de dinosaurio por cada metro cuadrado en la excavación que lleva a cabo. Elabore la distribución de probabilidad de Poisson con este dato y verifique si se cumple que $E(X) = Var(X) = np$.
14. Un comunicólogo estima que, en época electoral, aparecen en promedio 10 comerciales políticos cada media hora. Elabore la distribución de probabilidad de Poisson con este dato y verifique si se cumple que $E(X) = Var(X) = np$.
15. Un pescador estima que en promedio pesca 7 pescados grandes al día. Elabore la distribución de probabilidad de Poisson con este dato y verifique si se cumple que $E(X) = Var(X) = np$.

5.7 Distribución hipergeométrica

En la tabla 5.3 se resumieron las características que diferencian (y permiten distinguir) entre las 4 principales distribuciones discretas de probabilidad. Para facilitar la referencia se le reproduce en seguida:

Tabla 5.3 Características de las principales distribuciones discretas de probabilidad

Distribución	Núm. de resultados posibles	Probabilidad
Binomial	2 resultados posibles	Igual *P* en cada ensayo
Poisson	2 resultados posibles	Igual *P* en cada ensayo
Multinomial	Más de 2 resultados	Igual *P* en cada ensayo
Hipergeométrica	2 resultados posibles	*P* diferentes en cada ensayo

Tal como puede verse aquí, la distribución hipergeométrica se aplica cuando se tienen 2 resultados posibles y cuando la probabilidad, *P*, sí cambia de un ensayo a otro.

La fórmula de la distribución hipergeométrica de probabilidad es:

$$P(X) = \frac{\binom{N_1}{X}\binom{N_2}{n-X}}{\binom{N}{n}} = \frac{\left(\dfrac{N_1!}{X!(N_1 - X)!}\right)\left(\dfrac{N_2!}{(n-X)!(N_2 - n + X)!}\right)}{\dfrac{N!}{n!(N-n)}} \tag{5.9}$$

en donde:

N es el número total de elementos de la población.
N_1 es el número de elementos de interés en la población.
N_2 es el número de elementos contraparte de los de interés en la población.
n es el número de elementos de la muestra.
X es el valor de la variable aleatoria que interesa.

Se ilustra su uso en los ejemplos siguientes.

EJEMPLO 5.15

Una caja de 10 focos contiene 2 defectuosos (D) y 8 no defectuosos (N). Si se eligen al azar 3 focos de la caja, ¿cuál es la probabilidad de que la muestra contenga exactamente un foco defectuoso?

Solución:

En primer lugar, es importante observar que se cumplen las 2 características que hacen que este experimento aleatorio corresponda a una distribución hipergeométrica: se trata de una variable con 2 resultados posibles (D o N) y las probabilidades cambian en cada extracción: en la primera, la probabilidad de foco defectuoso es $\frac{2}{10} = 0.2$. La probabilidad de algún defectuoso en la segunda extracción cambia. Si el primer foco es defectuoso, la probabilidad de defectuoso en la segunda extracción es de $\frac{1}{9} = 0.11$ y si el primero es no defectuoso, esta probabilidad es de $\frac{2}{9} = 0.22$. La solución, en este caso, $N = 10$, $N_1 = 2$, $N_2 = 8$, $n = 3$ y $x = 1$, por lo que,

$$P(1) = \frac{\binom{N_1}{X}\binom{N_2}{n-X}}{\binom{N}{n}} = \frac{\binom{2}{1}\binom{8}{3-1}}{\binom{10}{3}}$$

$$= \frac{\binom{2}{1}\left(\frac{8 \times 7}{2}\right)}{\frac{10 \times 9 \times 8}{3 \times 2}} = \frac{2(28)}{120} = \frac{56}{120} = 0.46667$$

Con Excel: la sintaxis de la función hipergeométrica es:

DISTR.HIPERGEOM(muestra_éxito;núm_de_muestra;población_éxito;núm_de_población).

Y la función =DISTR.HIPERGEOM(1,3,2,10) da como resultado, precisamente, 0.46667.

EJEMPLO 5.16

De 10 obreros, 6 pertenecieron al sindicato por más de 7 años. Si de este grupo de 10 se eligen aleatoriamente 5 obreros, la probabilidad de que exactamente 3 de ellos tengan una antigüedad de más de 7 años es $N = 10$, $N_1 = 6$, $N_2 = 4$, $n = 5$ y $x = 3$, por lo que,

$$P(1) = \frac{\binom{N_1}{X}\binom{N_2}{n-X}}{\binom{N}{n}} = \frac{\binom{6}{3}\binom{4}{2}}{\binom{10}{5}}$$

$$= \frac{\left(\frac{6 \times 5 \times 4}{3 \times 2}\right)\left(\frac{4 \times 3}{2}\right)}{\frac{10 \times 9 \times 8 \times 7 \times 6}{5 \times 4 \times 3 \times 2}} = \frac{20(6)}{252} = \frac{120}{252} = 0.47619$$

Con Excel: =DISTR.HIPERGEOM(3,5,6,10) da como resultado, precisamente, 0.47619.

EJEMPLO 5.17

En una caja de 25 vasos, 7 tienen algún defecto. Si se eligen 10 vasos al azar:

a) ¿Cuál es la probabilidad de que en esa muestra se encuentren exactamente 3 vasos defectuosos?

b) ¿Cuál es la probabilidad de que 7 sean defectuosos?

Solución:

a) $N = 25$, $N_1 = 7$, $N_2 = 18$, $n = 10$ y $x = 3$,

$$P(3) = \frac{\begin{pmatrix} N_1 \\ X \end{pmatrix}\begin{pmatrix} N_2 \\ n - X \end{pmatrix}}{\begin{pmatrix} N_1 \\ X \end{pmatrix}} = \frac{\begin{pmatrix} 7 \\ 3 \end{pmatrix}\begin{pmatrix} 18 \\ 10 - 3 \end{pmatrix}}{\begin{pmatrix} 25 \\ 10 \end{pmatrix}}$$

$$= \frac{\left(\frac{5\,040}{144}\right)\left(\frac{18!}{7!11!}\right)}{\frac{25!}{10!15!}} = \frac{(35)(31\,824)}{3\,268\,760} = \frac{1\,113\,840}{3\,268\,760} = 0.3408$$

Con Excel:

=DISTR.HIPERGEOM(3,10,7,25) da como resultado, precisamente, 0.3408.

b) $N = 25$, $N_1 = 7$, $N_2 = 18$, $n = 10$ y $x = 7$,

$$P(7) = \frac{\begin{pmatrix} N_1 \\ X \end{pmatrix}\begin{pmatrix} N_2 \\ n - X \end{pmatrix}}{\begin{pmatrix} N \\ n \end{pmatrix}} = \frac{\begin{pmatrix} 7 \\ 7 \end{pmatrix}\begin{pmatrix} 18 \\ 10 - 7 \end{pmatrix}}{\begin{pmatrix} 25 \\ 10 \end{pmatrix}}$$

$$= \frac{(1)\left(\frac{18 \times 17 \times 16}{3 \times 2}\right)}{\frac{25!}{10!15!}} = \frac{816}{3\,268\,760} = 0.00025$$

Con Excel:

En primer lugar se verifica el número de combinaciones del denominador, que son 25 de 10 en 10. La función de Excel =COMBINAT(25,10) produce 3 268 760, con lo cual se asegura que ese complicado cálculo manual es correcto.

En seguida, la probabilidad hipergeométrica:

=DISTR.HIPERGEOM(7,10,7,25) da como resultado, precisamente, 0.00025.

5.7.1 Media y desviación estándar de la distribución hipergeométrica

La media aritmética y la varianza de una distribución hipergeométrica están dadas por las siguientes funciones:

$$= \frac{N_1 x}{N} \tag{5.10}$$

$$\sigma = \frac{nN_1(N - N_1)(N - n)}{N^2(N - 1)} \tag{5.11}$$

Para ilustrar estas propiedades de la distribución hipergeométrica se retoma el ejemplo 5.15 anterior como él.

Ejemplo 5.18

Se tenía una caja de 10 focos con 2 defectuosos (*D*) y 8 no defectuosos (*N*) y se eligieron al azar tres focos de la caja. En ese ejemplo se determinó que la probabilidad de que en esa muestra de tres focos hubiera 1 defectuoso era de 0.46667.

Ahora, construyendo la distribución de probabilidad completa para este experimento con una muestra de 3 focos, las probabilidades de obtener 0, 2 y 3 focos defectuosos son:

$$N_1 = 2;\ N_2 = 8;\ N = 10;\ n = 3$$

$$P(0) = \frac{\begin{pmatrix} N_1 \\ X \end{pmatrix}\begin{pmatrix} N_2 \\ n - X \end{pmatrix}}{\begin{pmatrix} N \\ n \end{pmatrix}} = \frac{\begin{pmatrix} 2 \\ 0 \end{pmatrix}\begin{pmatrix} 8 \\ 3 - 0 \end{pmatrix}}{\begin{pmatrix} 10 \\ 3 \end{pmatrix}}$$

$$= \frac{(1)\left(\frac{8 \times 7 \times 6}{3 \times 2}\right)}{\frac{10 \times 9 \times 8}{3 \times 2}} = \frac{56}{120} = \frac{56}{120} = 0.46667$$

$$P(2) = \frac{\binom{N_1}{X}\binom{N_2}{n-X}}{\binom{N}{n}} = \frac{\binom{2}{2}\binom{8}{3-2}}{\binom{10}{3}}$$

$$= \frac{(1)\left(\frac{8}{1}\right)}{\frac{10 \times 9 \times 8}{3 \times 2}} = \frac{8}{120} = \frac{56}{120} = 0.06667$$

$$P(3) = \frac{\binom{N_1}{X}\binom{N_2}{n-X}}{\binom{N}{n}}$$

$$= \frac{\binom{2}{3}\binom{8}{3-3}}{\binom{10}{3}} = 0$$

Esta última probabilidad es cero porque, como sólo hay 2 defectuosos en el lote de 10 focos, es imposible que haya 3 focos defectuosos en la muestra. Esto se nota también al sustituir los valores en la fórmula, ya que la expresión combinatoria $\binom{2}{3}$ no tiene sentido.

Con los resultados anteriores, la distribución de probabilidad para este experimento es:

Resultados (núm. de focos defectuosos) X	$P(X)$
0	0.46667
1	0.46667
2	0.06666
3	0
$\Sigma P(X_i)$	1*

* Se redondearon las cifras para asegurar que esta suma sea 1 y, así, evitar diferencias debidas a redondeo.

Se calculan en seguida la media y la varianza de esta distribución de probabilidad hipergeométrica:

X	$P(X)$	$X \cdot P(X)$	$X - \mu$	$(X - \mu)^2$	$P(X) \cdot (X - \mu)^2$
0	0.46667	0	–0.6	0.36	0.168001
1	0.46667	0.46667	0.4	0.16	0.074667
2	0.06666	0.13332	1.4	1.96	0.130654
3	0	0	2.4	5.76	0
$\Sigma P(X_i)$	1*	0.6			0.373322

Así,

$$\mu = 0.6, \text{ y}$$
$$\sigma^2 = 0.373322.$$

Ahora, de acuerdo con las fórmulas anotadas antes para la media y la varianza de una distribución hipergeométrica:

$$= \frac{N_1 x}{N} = \frac{3(2)}{10} = 0.6$$

$$\sigma^2 = \frac{nN_1(N - N_1)(N - n)}{N^2(N - 1)} = \frac{3(2)(10 - 2)(10 - 3)}{10^2(10 - 1)}$$

$$= \frac{6(8)(7)}{100(9)} = \frac{336}{900} = 0.3733$$

Como era de esperarse, son iguales a las calculadas con el procedimiento básico, sólo que también se puede observar que es más sencillo calcular estas medidas con estas últimas fórmulas.

EJEMPLO 5.19

En un estudio se preguntó a 15 personas qué salsa les gustaba más, 9 respondieron que la verde y 6 que la roja. Si se toma una muestra de 3:

a) Construya la distribución de probabilidad para las personas a las que les gusta más la salsa verde.
b) Calcule la media y la varianza a partir de esta distribución.
c) Calcule la media y la varianza a partir de las fórmulas.

Solución:

a)

X	$P(X)$	$X \cdot P(X)$	$X - \mu$	$(X - \mu)^2$	$P(X) \cdot (X - \mu)^2$
0	0.0440	0	–1.81	3.2761	0.1310
1	0.2967	0.2967	–0.81	0.6561	0.1968
2	0.4747	0.9495	0.19	0.0361	0.0170
3	0.1846	0.5538	1.19	1.4161	0.2691
	1	1.80			0.6139

$$\mu = 1.80$$
$$\sigma^2 = 0.6139$$

b) Con $N_1 = 9$; $N_2 = 6$; $N = 15$; $n = 3$:

$$= \frac{nN_1}{N} = \frac{3(9)}{15} = 1.8$$

$$\sigma^2 = \frac{nN_1(N - N_1)(N - n)}{N^2(N - 1)} = \frac{3(9)(15 - 9)(15 - 3)}{15^2(15 - 1)}$$

$$= \frac{27(6)(12)}{225(14)} = \frac{1\,944}{3\,150} = 0.6171$$

La pequeña diferencia entre los 2 valores encontrados para la varianza se debe a redondeo.

ejercicios 5.7 Distribución hipergeométrica

1. Cierto grupo se compone de 5 analistas y 9 técnicos. Si se eligen aleatoriamente a 5 personas para ser asignadas a un proyecto, ¿cuál es la probabilidad de que haya exactamente 2 analistas?

2. En un grupo de 20 estudiantes, 15 no están conformes con su calificación final. Si se pregunta a 4 de ellos:

a) ¿Cuál es la probabilidad de que exactamente 3 no estén conformes con su calificación?
b) ¿Cuál es la probabilidad de que al menos 3 no estén conformes con su calificación?

3. De un grupo de 10 trabajadores se seleccionan 3 para realizar un proyecto en la planta que se encuentra en Estados Unidos, si 4 trabajadores de los 10 ya fueron asignados a un proyecto anterior para ir a la misma planta, ¿cuál es la probabilidad que, de los 3 seleccionados,

a) ninguno haya estado en un proyecto con anterioridad?
b) uno haya estado en un proyecto con anterioridad?
c) dos hayan estado en un proyecto con anterioridad?
d) los tres hayan estado en un proyecto con anterioridad?

4. Una fábrica de muebles tiene 2 tiendas, una en el centro de la ciudad y otra en el norte. En la tienda del centro hay 40 empleados y 20 en la del sur. A 5 empleados se les aplicará un cuestionario sobre las condiciones en que laboran.

a) ¿Cuál es la probabilidad de que ninguno sea de la tienda del sur?
b) ¿Cuál es la probabilidad de que todos sean de la tienda del sur?

5. En un grupo de 25 niños hay 15 niñas y 10 niños; 5 faltaron el viernes pasado. ¿Cuál es la probabilidad de que:

a) dos de los ausentes sean niños?
b) dos de los ausentes fueran niñas?
c) todos fueran niños?
d) todos fueran niñas?

Media y desviación estándar de la distribución hipergeométrica

6. Con los datos del ejercicio 3, de un grupo de 10 trabajadores se selecciona a 3 para realizar un proyecto en la planta que se encuentra en Estados Unidos, si 4 trabajadores de los 10 ya fueron asignados a un proyecto anterior para ir a la misma planta.

a) Construya la distribución de probabilidad para los trabajadores asignados.
b) Calcule la media y la varianza a partir de esta distribución.
c) Obtenga la media y la varianza a partir de las fórmulas.

7. En un orfanato, 4 de cada 10 niños son menores de 3 años.

a) Elabore la distribución de probabilidades si se selecciona una muestra de 4 niños.
b) Calcule la media y la varianza a partir de esta distribución.
c) Obtenga la media y la varianza a partir de las fórmulas.

8. Después de la sangrienta batalla en el Alamein, un soldado del Africa Corps calcula que, de sus 15 amigos, 7 fueron capturados.

a) Construya la distribución de probabilidad completa para el número de soldados capturados en una muestra de 3.
b) Calcule la media y la varianza a partir de esta distribución.
c) Obtenga la media y la varianza a partir de las fórmulas.

9. Un oficial de La Armée de l'Air se enteró por un reporte de última hora que de los 25 aviones Dewoitine D.520 que contraatacaron a la Luftwaffe fueron destruidos 18.

a) Elabore la distribución de probabilidades si a cargo del oficial estaban 5 aviones.

b) Calcule la media y la varianza a partir de esta distribución.
c) Obtenga la media y la varianza a partir de las fórmulas.

10. De los 15 pacientes que atiende un psicoanalista, 5 sufren de esquizofrenia.

a) Elabore la distribución de probabilidades de los pacientes enfermos si se toma una muestra de 4 pacientes.
b) Calcule la media y la varianza a partir de esta distribución.
c) Obtenga la media y la varianza a partir de las fórmulas.

5.8 Distribución multinomial

La distribución multinominal se caracteriza por:

1. Más de 2 resultados posibles en cada ensayo.
2. Las probabilidades de los resultados no varían de un ensayo a otro; es decir, los diferentes ensayos son independientes.

La función de distribución de probabilidad multinomial es:

$$P(X_1, X_2, ..., X_k) = \frac{n!}{X_1!X_2!...X_k!} P_1^{x_1} P_2^{x_2} \cdots P_k^{x_k} \tag{5.12}$$

EJEMPLO 5.20

Los productos troquelados en una empresa son producidos por 4 trabajadores: González, Nava, Ruiz y López. Se sabe que, de los productos defectuosos, 30% son procesados por González, 25% por Nava, 40% por Ruiz y 5% por López. ¿Cuál es la probabilidad de que, entre 10 productos defectuosos elegidos al azar, 3 hayan sido elaborados por González, 2 por Nava, 3 por Ruiz y 2 por López?

Solución:

Aquí $n = 10$, $X_1 = 3$, $X_2 = 2$, $X_3 = 3$, $X_4 = 2$, $p_1 = 0.30$, $p_2 = 0.25$, $p_3 = 0.40$, $yp_4 = 0.05$.

Sustituyendo estos datos en la función:

$$P(X_1 = 3, \quad X_2 = 2, \quad X_3 = 3, X_4 = 2)$$

$$= \frac{10}{3!2!3!2!} (0.3)^3 (0.25)^2 (0.4)^3 (0.05)^2$$

$$= \frac{10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2}{3 \cdot 2 \cdot 2 \cdot 3 \cdot 2 \cdot 2} 0.027(0.0625)(0.064)(0.0025)$$

$$= 25\,200(0.0000003) = 0.006804$$

Lo que es igual que 0.68%, una probabilidad muy reducida pero que concuerda con el planteamiento, pues las probabilidades de ocurrencia en la muestra para los 4 eventos, 0.30, 0.20, 0.30 y 0.20, están un tanto distantes de las probabilidades que se dan en la población: 0.30, 0.25, 0.40 y 0.05.

EXCEL No existe función para esta distribución.

EJEMPLO 5.21

En un lote de 100 camisas, 50 son blancas, 30 son verdes y 20 son negras. Si se elige una muestra de 5 camisas, reemplazando cada vez que se extrae una, ¿cuál es la probabilidad de que en la muestra haya 2 camisas blancas, 2 verdes y una negra?

Solución:

En este caso, $X_1 = 2$, $X_2 = 2$, $X_3 = 1$, $p_1 = 0.50$, $p_2 = 0.30$, y $P_3 = 0.20$, por lo que

$$P(X_1, X_2, ..., X_k) = \frac{n!}{X_1!X_2!...X_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{xk}$$

$$= \frac{5!}{2!2!} 0.5^2(0.3)^2(0.2)$$

$$= \frac{5 \times 4 \times 3 \times 2}{2 \times 2}(0.25)(0.09)(0.2) = 0.135$$

5.8.1 Media y desviación estándar de la distribución multinomial

La distribución multinomial tiene tantas medias y varianzas como número de resultados posibles, y están dadas por las siguientes funciones:

$$E(X_i) = np_i \tag{5.13}$$

$$Var(X_i) = np_i\,(1 - p_i) \tag{5.14}$$

EJERCICIOS 5.8 Distribución multinomial

1. En una urna con 6 pelotas hay 3 rojas, 2 amarillas y una blanca. Si se saca al azar 5 veces una pelota, reemplazando cada vez, ¿cuál es la probabilidad de que sean 2 rojas, una amarilla y 2 blancas?
2. En una caja hay 3 monedas, una de cobre, una de plata y una de oro. ¿Cuál es la probabilidad de que al sacar 4 veces una moneda con reemplazo sea una de cobre, 2 de plata y una de oro?
3. En la biblioteca de una escuela primaria hay 200 libros: 30% de matemáticas, 10% de cuentos, 5% de idioma español, 20% de ciencias y 35% de historia. Si se eligen 15 libros al azar, ¿cuál es la probabilidad de que 2 sean de matemáticas, 3 de cuentos, 5 de español, 1 de ciencias y 4 de historia?
4. En una tienda trabajan 30 empleados: 7 cajeras , 5 cargadores, 15 vendedoras y 3 vigilantes, si se toma una muestra aleatoria de 8 trabajadores, ¿cuál es la probabilidad de que 4 sean cajeras, 1 cargador, 2 vendedoras y 1 vigilante?
5. En un lote con 20 autos hay: 2 morados, 2 verdes, 5 azules, 8 rojos y 3 blancos, se eligen 3 al azar. ¿Cuál es la probabilidad de que sean 1 verde, 1 rojo y 1 blanco?

Media y desviación estándar de una distribución multinomial

6. En una distribuidora de alimentos trabajan 30 empleados: 7 cajeras, 5 cargadores, 15 vendedoras y 3 vigilantes. Calcule:
   a) Las medias de la distribución.
   b) Sus varianzas.
7. En una bolsa de 30 paletas de sabores hay: 13 de fresa, 10 de limón, 7 de piña. Si se saca una muestra de 6 paletas, calcule:
   a) Las medias de la distribución.
   b) Sus varianzas.
8. En una empresa trabajan 100 empleados: 51 casados, 32 solteros, 13 divorciados y 4 viudos. Calcule:
   a) Las medias de la distribución.
   b) Sus varianzas.
9. En una escuela primaria se venden a los alumnos 3 tipos de bebidas en el recreo: gaseosas (G), agua embotellada (A) y jugos (J). La probabilidad de que un alumno elegido aleatoriamente compre en un día cualquiera una gaseosa es de 0.43, que compre una botella de agua de 0.20 y que compre un jugo es de 0.37. Calcule:
   a) Las medias de la distribución.
   b) Sus varianzas para un grupo de 200 alumnos.
10. Un general sabe que, durante una batalla, el riesgo de que el enemigo destruya una unidad móvil es alta y determinó que la probabilidad de que sea un avión (A) es 0.13, un helicóptero (H) 0.17, un tanque (T) 0.25 y un Jeep (J) 0.45. Calcule:
    a) Las medias de la distribución.
    b) Sus varianzas para un conjunto de 1 000 vehículos.

5.9 Resumen

En este capítulo se revisaron cuatro distribuciones de probabilidad para variables discretas, la binomial, la de Poisson, la hipergeométrica y la multinomial. Estas cuatro distribuciones se aplican a variables discretas, o discontinuas, que son variables que se distinguen sobre todo porque se distingue la diferencia entre un valor cualquiera de la variable y el anterior o el siguiente, cosa que no es posible con las variables continuas.

Se resumieron en un cuadro las principales características de estas distribuciones y qué las diferencia:

Distribución	Núm. de resultados posibles	Probabilidad
Binomial	2 resultados posibles	Igual p en cada ensayo
Poisson	2 resultados posibles	Igual p en cada ensayo
Multinomial	Más de 2 resultados	Igual p en cada ensayo
Hipergeométrica	2 resultados posibles	p diferentes en cada ensayo

En este cuadro se pueden distinguir claramente las diferencias entre todas ellas, excepto la que existe entre la binomial y la de Poisson. Y se explicó que la diferencia entre estas 2 es, en primer lugar, que la binomial trata de X éxitos en n ensayos, en tanto que la de Poisson aborda situaciones de promedios o valores esperados de éxitos en un continuo de tiempo o de espacio; en segundo lugar, se vio que, como la distribución de Poisson es el límite de la binomial cuando n es muy grande y p es muy pequeña, al darse estas circunstancias resulta más conveniente utilizar aquélla y no la binomial, pues es más fácil calcular la de Poisson e igual de precisa, aunque esto ya no es tan importante ahora, dada la amplia disponibilidad de computadoras.

Se vieron también los importantes conceptos de la media y la varianza de una distribución de probabilidad y se revisaron estas medidas para las cuatro distribuciones aquí tratadas; se señaló la especial importancia del concepto de "valor esperado de una variable" que es, de hecho, la media de su distribución de probabilidad. En capítulos posteriores se vuelve a este importante concepto.

5.10 Fórmulas del capítulo

5.3 Media y varianza de una distribución de probabilidades

Media de una distribución de probabilidad (valor esperado de la variable):

$$E(X) = \mu = \Sigma X_i \cdot P(X_i) \tag{5.1}$$

Varianza de la distribución de probabilidad:

$$Var(X) = \sigma^2 = \sum_{i=1}^{n} (X_i - \mu)^2 \cdot P(X_i) \tag{5.2}$$

5.4 Distribución binomial

$$q = 1 - p \tag{5.3}$$

$$P(x) = C_n^n p^x q^{n-x} = \frac{n!}{x!(n-x)!} p^x q^{n-x} \tag{5.4}$$

5.4.1 Media y varianza de la distribución binomial

Media aritmética:

$$\mu = np \tag{5.5}$$

Varianza:

$$\sigma^2 = npq \tag{5.6}$$

5.6 Distribución de Poisson

$$P(x) = \frac{e^{-\lambda}\lambda^x}{x!} \tag{5.7}$$

La media y la varianza de una distribución de Poisson:

$$E(X) = \mu = Var(X) = np \tag{5.8}$$

5.7 Distribución hipergeométrica

$$P(X) = \frac{\begin{pmatrix} N_1 \\ X \end{pmatrix}\begin{pmatrix} N_2 \\ n - X \end{pmatrix}}{\begin{pmatrix} N \\ n \end{pmatrix}}$$

$$= \frac{\left(\dfrac{N_1!}{X!(N_1 - X)!}\right)\left(\dfrac{N_1!}{(n - X)!(N_2 - n + X)!}\right)}{\dfrac{N!}{n!(N - n)!}} \tag{5.9}$$

5.7.1 Media y desviación estándar de la distribución hipergeométrica

$$= \frac{N_1 x}{N} \tag{5.10}$$

$$\sigma^2 = \frac{nN_1(N - N_1)(N - n)}{N^2(N - 1)} \tag{5.11}$$

5.7 Distribución multinomial

$$P(X_1, X_2, ..., X_k) = \frac{n!}{X_1!X_2!...X_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k} \tag{5.12}$$

5.8.1 Media y desviación estándar de la distribución multinomial

$$E(X_i) = np_i \tag{5.13}$$

$$Var(X_i) = np_i(1 - p_i) \tag{5.14}$$

5.11 Ejercicios adicionales

5.3 Media y varianza de una distribución de probabilidades

1. En un año, el Índice de Precios y Cotizaciones de la Bolsa Mexicana de Valores operó 256 días y durante 154 tuvo precios de cierre al alza. En la tabla siguiente se muestra el número de cierres al alza durante cada uno de los 12 meses de ese año.

Meses	Alzas
Enero	13
Febrero	13
Marzo	13
Abril	10
Mayo	15
Junio	11
Julio	9
Agosto	13
Septiembre	15
Octubre	13
Noviembre	13
Diciembre	16
	154

Determine:

a) Las probabilidades asociadas a cada evento.
b) La media de la distribución de probabilidad.
c) Su varianza.

2. En la tabla siguiente se muestra el número de autos que se reparan en un taller mecánico durante un periodo de 50 días.

Autos	0	1	2	3	4	5	6	Suma
Días	4	5	8	13	12	3	5	50

a) Determine las probabilidades asociadas a cada evento.
b) Calcule la media de la distribución de probabilidad.
c) Obtenga su varianza.

3. Un emprendedor pretende exportar artesanías de barro negro a Alemania e investigó cuántas figuras de barro produce diariamente un artesano en un periodo de 25 días. En la tabla siguiente se muestran los resultados de la producción diaria de figuras de barro negro de un artesano.

Artesanías	Días
10	2
11	3
12	5
13	8
14	7
Suma	25

a) Determine las probabilidades asociadas a cada evento.
b) Calcule la media de la distribución de probabilidades.
c) Obtenga su varianza.

4. El número de casas de interés social que vende una empresa hipotecaria por semana sigue la distribución de probabilidad de la tabla siguiente.

Núm. de casas	0	1	2	3	4	5	6
	0.03	0.06	0.19	0.32	0.28	0.07	0.05

a) Calcule la media de la distribución de probabilidades.
b) Obtenga su varianza.

5. En un hospital se sabe que el número de niños que entra cada hora a urgencias sigue la distribución de probabilidad de la tabla siguiente.

Núm. de niños	0	1	2	3	4	5
$P(X_i)$	.08	.12	.30	.30	.12	.08

a) Calcule la media de la distribución de probabilidades.
b) Obtenga su varianza.

6. En una fábrica de planchas se sabe que el número de productos que son terminados cada hora siguen la distribución de probabilidad de la tabla siguiente.

Núm. de planchas	50	51	52	53	54	55
$P(X)$	0.06	0.16	0.28	0.32	0.13	0.05

a) Calcule la media de la distribución de probabilidades.
b) Obtenga su varianza.

7. Para coordinar sus entregas con sus clientes, una fábrica determinó el número de productos que elabora en 100 días. En la tabla siguiente se muestra el número de productos terminados durante ese periodo de 100 días.

Ventas	20	21	22	23	24	25	Suma
Días	10	18	22	21	19	10	100

a) Determine las probabilidades asociadas a cada evento.
b) Calcule la media de la distribución de probabilidades.
c) Obtenga su varianza.

8. Para estimar el número de bajas en la próxima batalla, un general determinó cómo se distribuyeron durante las últimas 12 batallas. En la tabla siguiente se muestra el número de bajas durante esas últimas batallas.

Bajas	56	57	58	59	60	Suma
Batalla	1	3	5	2	1	12

a) Determine las probabilidades asociadas a cada evento.
b) Calcule la media de la distribución de probabilidades.
c) Obtenga su varianza.

9. En la tabla siguiente se muestra el número de remesas que recibió un municipio en un estado del norte del país durante un periodo de 60 días.

Remesas	1	2	3	4	5	6	7	8	9	10	Suma
Días	2	3	5	7	8	12	9	6	5	3	60

a) Determine las probabilidades asociadas a cada evento.
b) Calcule la media de la distribución de probabilidades.
c) Obtenga su varianza.

5.4 Distribución binomial

10. Diariamente salen 10 vuelos del Aeropuerto Internacional de la Ciudad de México hacia el de Guadalajara. La probabilidad de que un vuelo se retrase es de 25 por ciento.

a) ¿Cuál es la probabilidad de que ningún vuelo se retrase?
b) ¿Cuál es la probabilidad de que un vuelo se retrase?

11. La probabilidad de sufrir un accidente en la carretera es de 22%, si se sigue la trayectoria de 10 autos:

a) ¿Cuál es la probabilidad de que ninguno sufra un accidente?
b) ¿Cuál es la probabilidad de que la mitad sufra algún accidente?
c) ¿Cuál es la probabilidad de que todos sufran un accidente?

12. Suponiendo que 60% de todas las personas prefiere el helado de chocolate a otros sabores, y se seleccionan 20 personas para un estudio, ¿cuál es la probabilidad de que entre 5 y 10 prefieran el helado de chocolate?

13. La probabilidad de que una persona que solicita una tarjeta de crédito se considere sujeto de crédito es de 0.43. Si la solicitan 11 personas, ¿cuál es la probabilidad de que exactamente 5 sean sujetos de crédito?

14. En relación con el problema anterior, ¿cuál es la probabilidad de que 6 o menos personas sean sujetos de crédito?

15. Si en un día 22 personas solicitan una tarjeta de crédito y la probabilidad de que una persona sea sujeto de crédito es de 0.43, ¿cuál es la probabilidad de que 7 lo sean?

16. La probabilidad de que una acción de la Bolsa Mexicana de Valores aleatoriamente elegida reparta dividendos en efectivo este año es de 0.25. Si se elige aleatoriamente a 12 acciones:

a) ¿Cuál es la probabilidad de que 4 de ellas repartan utilidades en efectivo?
b) ¿Cuál es la media de esta distribución de probabilidad?
c) ¿Cuál es la varianza de esta distribución de probabilidad?

17. La probabilidad de que una acción de la Bolsa Mexicana de Valores elegida al azar reparta dividendos en efectivo este año es de 0.23. Si se eligen aleatoriamente 8 acciones:

a) ¿Cuál es la probabilidad de que todas ellas repartan utilidades en efectivo?
b) ¿Cuál es la probabilidad de que repartan utilidades en efectivo más de 5 de ellas?
c) ¿Cuál es la media de esta distribución de probabilidad?
d) ¿Cuál es la varianza de esta distribución de probabilidad?

5.4.1 Media y varianza de la distribución binomial

18. La probabilidad de que este año una secretaría de Estado, seleccionada aleatoriamente, cumpla con las expectativas de reducción de gastos es de 0.67.

a) Calcule el espacio muestral de probabilidades de que las secretarías de una muestra de 4 cumplan la expectativa.
b) Obtenga la media de la distribución de probabilidad.
c) Determine su varianza por los 2 métodos a partir de la distribución de probabilidad y del de la fórmula para la varianza.

19. La probabilidad de que un accionista seleccionado aleatoriamente vote en favor de la fusión propuesta por la alta dirección es de 0.51.

a) Calcule el espacio muestral de probabilidades de que 4 accionistas voten en favor.
b) Obtenga la media de la distribución de probabilidad.
c) Determine su varianza por los 2 métodos, a partir de la distribución de probabilidad y del de la fórmula para la varianza.

5.5 Tres formas de presentar una distribución de probabilidad

20. Se sabe que la probabilidad de que un foco de una mampara se funda es de 8%. Calcule el espacio muestral de que se funda un foco en una muestra de 7 mamparas y preséntelo en forma de:

a) Función.
b) Tabla.
c) Gráfica de barras.

21. De las cuerdas vendidas por un pequeño productor, 5.5% de ellas se rompen por algún defecto. Calcule el espacio muestral de que se rompan cuerdas en una muestra de 3 y preséntela en forma de:

a) Función.
b) Tabla.
c) Gráfica de barras.

22. La probabilidad de que una botella de muestra que se elabora para el lanzamiento de un perfume se rompa en el trayecto de la planta al punto de venta es de 12%. Calcule el espacio muestral de que se rompan botellas en una muestra de 5 y preséntela en forma de:

a) Función.
b) Tabla.
c) Gráfica de barras.

23. Treinta y siete por ciento de los clientes que entran a una sucursal bancaria realizan pagos de algún tipo de tarjeta de crédito. Calcule el espacio muestral de una muestra de 4 para el pago de alguna tarjeta de crédito y preséntela en forma de:

a) Función.
b) Tabla.
c) Gráfica de barras.

24. El departamento de salud industrial informó que 41% de los trabajadores de una empresa sufren de estrés laboral. Calcule el espacio muestral de que trabajadores sufran estrés en una muestra de 6 y preséntela en forma de:

a) Función.
b) Tabla.
c) Gráfica de barras.

5.5 Distribución de Poisson

25. A un muelle llega en promedio un barco cada 2 días. ¿Cuál es la probabilidad de que lleguen en un día 2 o más barcos?

26. El borrador de un libro tiene un total de 100 errores en las 1 000 páginas que lo conforma. ¿Cuál es la probabilidad de que:

a) una sección de 30 páginas tenga 2 o más errores?
b) una sección de 50 páginas tenga 2 o más errores?
c) una página elegida al azar no tenga ningún error?

27. En promedio 6 personas por hora hacen alguna transacción en un cajero electrónico ubicado en un supermercado. ¿Cuál es la probabilidad de que:

a) exactamente 6 usen el cajero durante una hora aleatoriamente seleccionada?
b) ninguna persona lo utilice durante un periodo de 10 minutos?
c) ninguna persona lo utilice durante un periodo de 5 minutos?

28. Una empresa planea ampliar la cobertura del seguro de gastos médicos de sus 3 000 trabajadores para una enfermedad extraña. Si la probabilidad de que una persona tenga este padecimiento es de 0.001:

a) ¿Cuál es el número esperado de trabajadores que tendrán esta enfermedad?
b) ¿Cuál es la probabilidad de que ninguno de los 3 000 trabajadores presente la enfermedad?

29. La probabilidad de que una máquina fabrique tuercas defectuosas es de 0.06. Si se produjeron 450 tuercas en el primer turno:

a) ¿Cuál es el número esperado de tuercas defectuosas?
b) ¿Cuál sería el número esperado de tuercas defectuosas si la probabilidad se duplica?

5.6.2 Distribución de Poisson como aproximación de la distribución binominal

30. Según un informe publicado por una universidad, la probabilidad de que un estudiante con promedio de 9.0 al terminar el bachillerato no ingrese en su primera opción de la licenciatura es de $p = 0.004$. Determine la probabilidad de que de 1 500 alumnos con un promedio de 9.0, cuando mucho 3 no ingresen en su primera opción.

a) Use la distribución de Poisson.
b) Utilice la binomial.

31. El servidor de internet de una empresa presenta algún tipo de falla 7% del tiempo. Si se revisa en 200 ocasiones, ¿cuál es la probabilidad de que en 3 exista algún tipo de falla? Utilice:

a) La distribución de Poisson.
b) La binomial.

32. Un supermercado recibe un embarque de manzanas. Si la probabilidad de que una pieza esté maltratada es de 1%, determine la probabilidad de que, al elegir 350 al azar, menos de 3 se maltrataron usando:

a) La distribución de Poisson.
b) La binomial.

33. La probabilidad de que un bebé tenga algún tipo de daño grave si sufre un golpe en la cabeza es de 0.006. Si se toman 280 casos atendidos en un hospital, ¿cuál es la probabilidad de que ninguno presente daño alguno? Utilice:

a) La distribución de Poisson.
b) La binomial.

34. Un fabricante de zapatos determinó que la probabilidad de que una suela se despegue durante el primer año de uso es de 2.5%. Si se seleccionan 350 piezas elaboradas en el último año, determine la probabilidad de que al menos 5 se les despegara la suela en el primer año usando:

a) La distribución de Poisson.
b) La binomial.

5.6.3 Media y varianza de la distribución de Poisson

35. Se realizó un estudio para conocer el número de accidentes de trabajo que se presentaban en las diferentes plantas de una empresa textilera. En la siguiente tabla se muestra la posibilidad de que se presente algún accidente en el transcurso de un mes.

a) Calcule el número esperado de accidentes por mes.
b) Calcule la varianza.

Núm. de accidentes	$P(X_i)$
0	0.1353
1	0.2707
2	0.2707

(*continúa*)

(*continuación*)

Núm. de accidentes	$P(X_i)$
3	0.1804
4	0.0902
5	0.0361
6	0.0120
7	0.0034
8	0.0009
9	0.0002
Total	0.9999

36. Un vendedor de verduras en un mercado ambulante recibe la mercancía en cajas. En la siguiente tabla se muestra la posibilidad de que encuentre alguna verdura podrida en alguna de ellas.

a) Calcule el número esperado de verduras podridas por caja.
b) Calcule la varianza.

Verduras podridas X_i	$P(X_i)$
0	0.0821
1	0.2052
2	0.2565
3	0.2138
4	0.1336
5	0.0668
6	0.0278
7	0.0099
8	0.0031
9	0.0009
10	0.0002
Total	0.9999

37. Una máquina fabrica envases para agua y jugo. En la siguiente tabla se muestra la probabilidad de que algún envase presente cualquier tipo de fuga.

a) Calcule el número esperado de fugas.
b) Calcule la varianza.

Verduras podridas X_i	$P(X_i)$
0	0.2231
1	0.3347
2	0.2510
3	0.1255
4	0.0471
5	0.0141
6	0.0035
7	0.0008
8	0.0001
Total	0.9999

38. Una tienda de artículos de computación recibe lotes de discos compactos cada semana de parte de su proveedor. A continuación se muestra la tabla con la posibilidad de encontrar discos con defectos de fábrica.

a) Calcule el número esperado de discos con defectos.
b) Calcule la varianza.

Discos con defectos X_i	$P(X_i)$
0	0.02
1	0.05
2	0.11
3	0.17
4	0.26
5	0.24
6	0.09
7	0.05
8	0.01
Total	1.00

39. La comisión de energía eléctrica realiza un estudio para conocer el número de casas que obtienen el servicio de manera irregular mediante las conexiones ilegales conocidas como "diablitos". A continuación se muestra la probabilidad de encontrar "diablitos" en una calle.

a) Calcule el número esperado de diablitos.
b) Calcule la varianza.

"Diablitos" por calle X_i	$P(X_i)$
0	0.01
1	0.06
2	0.12
3	0.17
4	0.24
5	0.14
6	0.09
7	0.07
8	0.05
9	0.03
10	0.02
Total	1

5.7 Distribución hipergeométrica

40. De 10 obreros, 6 pertenecieron al sindicato por más de 7 años. Si de este grupo de 10 se eligen aleatoriamente 5 obreros, ¿cuál es la probabilidad de que exactamente 3 de ellos tengan una antigüedad de más de 7 años?

41. De 25 candidatos a un puesto, 15 tienen experiencia laboral. Si de este grupo se eligen aleatoriamente a 8 candidatos, ¿cuál es la probabilidad de elegir exactamente 2 que tengan experiencia laboral?

42. De 12 alumnos que presentan su examen profesional a final de mes, 4 son candidatos a mención honorífica. Si de este grupo se eligen aleatoriamente a 8 alumnos, ¿cuál es la probabilidad de elegir exactamente 2 que sean candidatos a mención honorífica?

43. Un ingeniero encargado del área de nuevas tecnologías de una empresa automotriz sabe que, de cada 10 proyectos que se presentan a la alta dirección, se autorizan solamente 3. Si está en espera de que le autoricen 5 nuevos proyectos, ¿cuál es la probabilidad de que le sean aprobados 2?

44. En un camión de la policía se transportan 25 reclusos, de los cuales 13 están presos por primera vez. Si se seleccionan al azar a 7, ¿cuál es la probabilidad de que 5 estuvieron presos en otras ocasiones?

5.7.1 Media y desviación estándar de la distribución hipergeométrica

45. Durante los últimos 3 años de su gobierno, un presidente municipal promovió 11 proyectos de desarrollo sustentable, de los cuales 6 fueron de educación primaria y secundaria.

a) Elabore la distribución de probabilidades para los proyectos de desarrollo sustentable si desea seleccionar 3 para un discurso.
b) Calcule la media y la varianza a partir de esta distribución.
c) Obtenga la media y la varianza a partir de las fórmulas.

46. En un frasco se colocaron 20 pastillas, de las cuales 8 son de azúcar. Si se toman 4 pastillas al azar:

a) Construya la distribución de probabilidad para las pastillas de azúcar.
b) Calcule la media y la varianza a partir de esta distribución.
c) Obtenga la media y la varianza a partir de las fórmulas.

47. Se recibe un embarque de granadas y, en una caja con 30 piezas, hay 5 que no detonarán al momento de usarlas. Si se toma una muestra aleatoria de 3 granadas:

a) Construya la distribución de probabilidad para las granadas que no detonarán.
b) Calcule la media y la varianza a partir de esta distribución.
c) Obtenga la media y la varianza a partir de las fórmulas.

48. A un bar llega un grupo de 18 jóvenes, de los cuales 8 tienen la mayoría de edad. Si la mesera les pide su identificación a sólo 6 de ellos:

a) Construya la distribución de probabilidad para los que no tienen la mayoría de edad.
b) Calcule la media y la varianza a partir de esta distribución.
c) Obtenga la media y la varianza a partir de las fórmulas.

49. Se tiene a un grupo de 15 personas diagnosticadas con diabetes: 8 de ellas tienen diabetes tipo 1. Se elige a 5 de ellos.

a) Construya la distribución de probabilidad para los diabéticos tipo 1.
b) Calcule la media y la varianza a partir de esta distribución.
c) Obtenga la media y la varianza a partir de las fórmulas.

5.8 Distribución multinomial

50. Del total de manuscritos elaborados por un grupo secretarial, 20% son elaborados por María, 25% por Rosa, 15% por Martha y 40% por Leticia. Si se toman 4 escritos, ¿cuál es la probabilidad de encontrar un manuscrito por secretaria?

51. Debido al fuerte y creciente presupuesto destinado a preparar sus atletas, un país estima que las probabilidades de obtener medallas de oro en natación (A), halterofilia (B), judo (C) y gimnasia olímpica (D) son de 0.3, 0.15, 0.21 y 0.32, respectivamente. Si dentro de esas disciplinas el país ganó 5 medallas de oro, ¿cuál es la probabilidad de que fueron: una en natación, una en halterofilia, una en judo y 2 en gimnasia?

52. Un informe del departamento de mercadotecnia señala las probabilidades de que se rechace un proyecto de inversión por los siguientes motivos: *a*) porque el estudio de mercado considera que no hay demanda suficiente para el nuevo producto (0.18); *b*) es inviable porque el estudio técnico así lo indica (0.22); *c*) porque el estudio financiero determina que no es rentable (0.60). Si este año se rechazaron 7 proyectos de inversión, ¿cuál es la probabilidad de que fueron rechazados 2 por el estudio de mercado, 2 por el estudio técnico y 3 por el estudio financiero?

53. Cuando el ejecutivo federal propone una nueva iniciativa de ley, la probabilidad de que un diputado vote por ella y que sea del partido *A*, partido *B* o partido *C*, es del 0.2, 0.3 y 0.5, respectivamente. Si se eligen, aleatoriamente, 8 diputados que votaron en favor de una incitativa de ley propuesta por el presidente, ¿cuál es la probabilidad de que 2 pertenezcan al partido *A*, 2 al partido *B* y 4 al partido *C*?

54. Estadísticas sobre población rural señalan que la probabilidad de que, al crecer, un joven se haga agricultor y permanezca en el campo (*A*) es de 0.18; de que se dedique a cualquier otro oficio dentro de su lugar de origen (*B*) es de 0.25; de que emigre a una ciudad (*C*) es de 0.21, y de que emigre a Estados Unidos (*D*) es de 0.23. Si se tomaran al azar 15 expedientes provenientes de escuelas rurales de enseñanza básica de ex alumnos que estudiaron 15 años atrás, ¿cuál es la probabilidad de que hoy 5 de ellos sean agricultores, 4 tengan un oficio diferente en su lugar de origen, 3 vivan en una ciudad y 3 emigraron a Estados Unidos?

5.8.1 Media y desviación estándar de la distribución multinomial

55. Se organizó una convención en una playa de la República Mexicana. La probabilidad de que un participante llegue en avión es de 0.46, en autobús 0.24 y en auto propio de 0.3. Si se elige a un grupo de 7 personas, calcule las medias para:

a) Dos que llegarán en avión.
b) Tres en autobús.
c) Dos en auto propio.
d) Calcule las varianzas correspondientes.

56. Existen 4 partidos políticos que participarán en las próximas elecciones, la probabilidad de que una persona vote por el partido rojo es de 0.26, por el verde 0.13, por el amarillo 0.38 y por el negro 0.23. Si se les pregunta a 12 personas, calcule las medias para:

a) Tres que votarán por el partido rojo.
b) Dos por el verde.
c) Uno por el amarillo.
d) Seis por el negro.
e) Calcule las varianzas correspondientes.

57. Se sabe que la camada de la pareja de un perro negro y uno blanco tiene la probabilidad de que 0.41 sean negros, 0.32 blancos, 0.19 pintos y 0.08 de algún otro color. Si se tiene una camada de 8 cachorros, calcule las medias para:

a) Tres negros.
b) Dos blancos.
c) Un pinto.
d) Dos de otro color.
e) Calcule las varianzas correspondientes.

58. En una bolsa hay dulces de varios tipos. Si se toma 1 al azar, la probabilidad de que salga un tamarindo es de 0.11, 0.23 de que sea paleta, 0.17 de que sea caramelo, 0.19 de que sea chicle, 0.14 de que sea chocolate y 0.16 de que sea bombón. Si se sacan 12 dulces al azar, calcule las medias para:

a) Un tamarindo.
b) Cuatro paletas.
c) Dos caramelos.
d) Un chicle.
e) Dos chocolates.
f) Tres bombones.
g) Calcule las varianzas correspondientes.

59. En un grupo de estudiantes se sabe que la probabilidad de que cualquiera de ellos naciera en el mes de enero es de 0.2, febrero 0.15, marzo 0.12, abril 0.23, mayo 0.19 y junio 0.11. Se eligen 30 personas al azar, calcule entonces las medias para:

a) Cuatro estudiantes nacidos en enero.
b) Dos en febrero.
c) Tres en marzo.
d) Seis en abril.
e) Siete en mayo.
f) Ocho en junio.
g) Calcule las varianzas correspondientes.

capítulo 6

Distribuciones continuas de probabilidad

Sumario

6.1 Área como medida de probabilidad
6.2 Distribución normal de probabilidad
6.2.1 Características de la distribución normal
6.2.2 Distribución normal estándar
6.2.3 Tabla de áreas bajo la curva normal
6.2.4 Determinación de probabilidades para cualquier distribución normal
6.3 Ajuste cuando se utiliza la distribución normal para evaluar probabilidades de una variable discreta (ajuste por discontinuidad)
6.4 Aproximación de distribuciones de probabilidad de variables discontinuas con la distribución normal
6.4.1 Aproximación de la distribución binomial con la distribución normal
6.4.2 Aproximación de la distribución de Poisson con la distribución normal
6.5 Distribución exponencial de probabilidad
6.5.1 Relación entre la distribución exponencial y la distribución de Poisson
6.6 Otras distribuciones de probabilidad continuas
6.7 Advertencia
6.8 Resumen
6.9 Fórmulas del capítulo
6.10 Ejercicios adicionales

Las distribuciones que se revisaron hasta aquí eran aplicables para variables aleatorias discretas, es decir, era posible contar los resultados posibles, y cada uno de ellos tenía una probabilidad identificable que, por lo general, era diferente de 0. Además, también es lo más común que el número de estos resultados posibles sea finito y, más aún, reducido (como 2 en el caso de la binomial).

Por otro lado, las características que distinguen a las distribuciones continuas es que los resultados posibles no se obtienen contando, sino midiendo, lo cual hace posible obtener observaciones tan precisas que se pierde la diferencia entre observaciones consecutivas. El único impedimento para lograr mediciones extremadamente precisas es el instrumento con que se toman, pero la posibilidad de lograrlo persiste. Así, como en una línea continua existe un número infinito de puntos, no es posible determinar la probabilidad de que se presente un valor específico, es decir, la probabilidad de ocurrencia de cada uno de los infinitos resultados posibles es 0, por lo que es necesario manejar la probabilidad de que una medición determinada se encuentre dentro de un intervalo de valores.

A diferencia de las distribuciones discretas, cuyas gráficas se presentan como diagramas de barras, las gráficas de las distribuciones continuas son líneas suavizadas, como las de la campana de la distribución normal, que es la más importante de todas las distribuciones de probabilidad, tanto discretas como continuas y que es el principal tema de este capítulo.

Se analiza también aquí otra importante distribución continua de probabilidad, la distribución exponencial.

6.1 Área como medida de probabilidad

Esa probabilidad medida con intervalos se puede ilustrar mediante la superficie de un cuadrado, como el que se muestra en la figura 6.1. Esa figura tiene una superficie de 100 unidades cuadradas, ya que tiene 10 de altura y 10 de lado. Y la porción de superficie rayada entre las 2 líneas horizontales (el eje horizontal de la gráfica y la línea trazada a la altura del 10) y entre las líneas verticales trazadas sobre los puntos 6 y 9 es 30% del área total. Esto mismo, interpretado en términos de probabilidad, permite afirmar que si se escoge al azar cualquier punto de ese cuadrado, existe una probabilidad de 30% de que se encuentre entre 6 y 9 (y dentro de las 2 líneas horizontales).

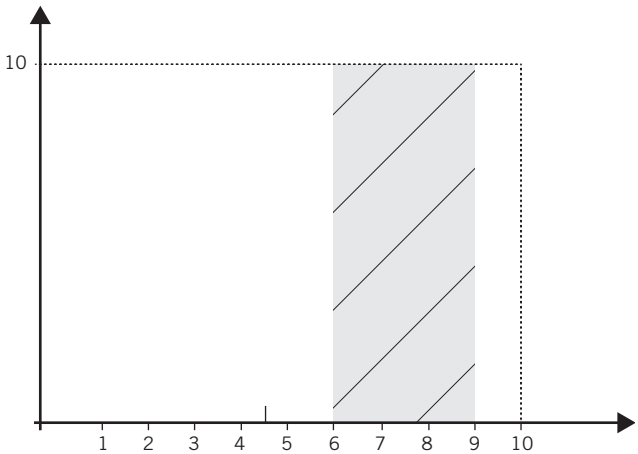


Figura 6.1 Un cuadrado con superficie de 100 unidades cuadradas.

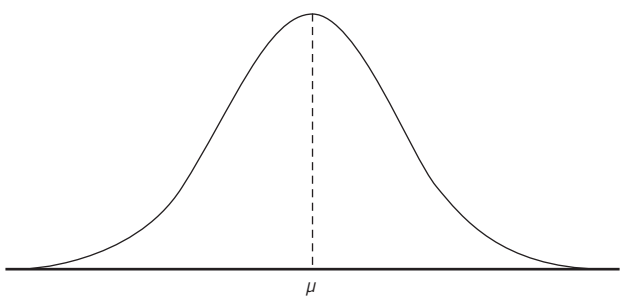


Figura 6.2 Una curva normal.

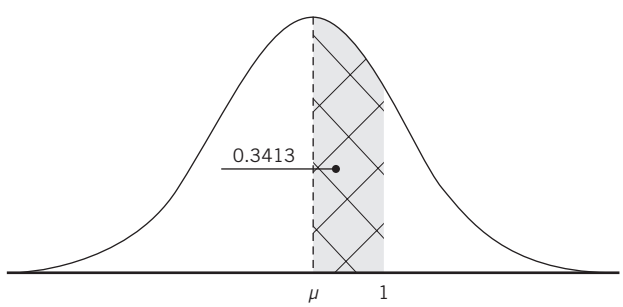


Figura 6.3 La porción de área que se encuentra entre la media y una desviación estándar de una distribución normal.

En otras palabras, la sección achurada en la figura 6.1 representa 30% del total de la superficie del cuadrado y la interpretación convierte esta proporción de superficie en probabilidad y lo mismo se aplica para interpretar probabilidades para conjuntos de datos que tienen forma de campana (distribución normal), como se explica y ejemplifica en las secciones siguientes.

6.2 Distribución normal de probabilidad

En las distribuciones continuas de probabilidad, como la normal, las probabilidades se representan en la misma forma gráfica, como porciones de área que se encuentran entre 2 líneas verticales, por encima del eje horizontal y por debajo de la gráfica de la función de la distribución. La distribución normal que es, con mucho, la más importante de las distribuciones continuas, tiene forma de campana, como se muestra en la figura 6.2.

Como se verá más adelante a detalle, la superficie que se encuentra entre las 2 rayas verticales trazadas sobre la media y una desviación estándar, por debajo de la campana normal, y por encima del eje de las y es 0.3413 o 34.13%, que es el área que se ilustra en la figura 6.3.

La función de densidad de esta distribución fue publicada por primera vez por Abraham de Moivre (1667-1754) en 1733 y la desarrolló como una aproximación de la distribución binomial; posteriormente Pierre Simon, Marqués de Laplace (1749-1827), la usó en 1783 para estudiar errores de medición; en 1809, Kart Friederich Gauss (1777-1855) la usó en el análisis de datos astronómicos.

6.2.1 Características de la distribución normal

La forma gráfica de esta distribución, en forma de campana, implica varias propiedades:

1. En primer lugar, es fácil notar que se trata de una figura simétrica y el eje de simetría es precisamente la media de la distribución, marcada como μ y como la línea punteada de las figuras 6.2 y 6.3.
2. Los valores de la variable hacia ambos extremos, los valores positivos y los negativos, se extienden hasta el infinito. En símbolos: $-\infty < x < \infty$. En términos gráficos, los extremos de esta curva nunca tocan el eje horizontal: son asintóticos.
3. La simetría de la curva respecto a su punto medio implica que la media, la mediana y la moda son iguales: $\mu = Md = Mo$.
4. La forma de la distribución normal también implica que la mayor parte de las observaciones están cerca del centro (de ahí que la parte más alta de la curva esté precisamente en medio), y se aplanan entre más alejados del centro se encuentren los puntos en ambos sentidos, muchas variables aleatorias tienen estas características.

Algunos ejemplos de las numerosas variables aleatorias que tienen la característica de agruparse mayoritariamente alrededor de la media son mediciones en seres vivos, como la estatura: la mayor parte de las personas y de los animales en general son medianos; hay algunos altos y muy pocos muy altos, de la misma manera que hay algunos bajos de estatura y muy pocos, muy bajos. Lo mismo sucede con otras variables como el peso. Otras ilustraciones se dan en el caso de procesos de producción; por ejemplo, en el llenado de botellas de refresco que contienen más o menos medio litro, es común que la mayor parte de las botellas

tengan un contenido cercano al medio litro y que haya pocas que tengan 505 o 495 mililitros. Lo mismo ocurre con alimentos enlatados.

Otros ejemplos de variables que suelen tener una distribución normal o aproximadamente normal: la longitud o peso de piezas metálicas troqueladas, la duración de llantas o focos en condiciones similares, y muchos más.

6.2.2 Distribución normal estándar

Tal como se vio antes, la función de probabilidad para variables aleatorias discretas, como la binomial o la de Poisson, calculan la probabilidad de ocurrencia de un resultado posible específico, mientras que, para variables continuas como la normal, se tiene una *función de densidad de probabilidad* que para la normal es:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \tag{6.1}$$

En donde e y π son las conocidas constantes cuyos valores aproximados son 2.71828 y 3.1416, respectivamente, σ es la desviación estándar de la distribución y μ es su media o promedio aritmético. Como la cantidad posible de combinaciones de valores de estas 2 medidas es infinita, se tiene entonces que hay una cantidad infinita de distribuciones normales, cada una de ellas definida por el valor específico de su media aritmética y de su desviación estándar.

Si se dan valores específicos a la desviación estándar y a la media y se calculan valores de $f(x)$ para diversos valores de la variable aleatoria X, y se dibujan estos resultados, entonces se obtiene la figura de campana que se muestra en las figuras 6.2 y 6.3.

Además, si se fijan valores de $\mu = 0$ y $\sigma = 1$, entonces la función de densidad de la distribución normal se simplifica a:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(X-\mu)^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi}} e^{\frac{-X^2}{2}}$$

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-X^2}{2}} \tag{6.2}$$

Esta última expresión es la función de densidad de la distribución normal estándar, la cual se caracteriza porque, además de ser normal, su media es igual a 0 y su desviación estándar es de 1. En símbolos: $\mu = 0$ y $\sigma = 1$.

Con esta función de densidad se pueden calcular probabilidades de áreas bajo la curva normal, y sobre el eje x, entre 2 valores (un intervalo) determinado de X. Para hacer esto se requiere utilizar procedimientos de cálculo integral fuera del alcance de este texto; sin embargo, existen valores tabulados de estas áreas, que se incluyen en este libro para que sea fácil consultarlas y cuya utilización se explica en la sección siguiente. También se pueden obtener estos valores mediante 4 funciones de Excel, las cuales se explican e ilustran en la sección 6.2.3.1. También se puede construir una tabla de áreas bajo la curva normal como la que se reproduce en este libro y en la tabla 6.1 y que es la que se suele utilizar para determinar estas áreas (cuando no se hace directamente con Excel o con algún otro mecanismo computarizado).

6.2.3 Tabla de áreas bajo la curva normal

Esta tabla permite encontrar cualquier área bajo la curva normal estándar delimitada entre 2 puntos. Aquí, es necesario recordar que esta distribución normal estándar se distingue porque su media es igual a 0 y su desviación estándar es igual a 1 y, además, como la distribución es simétrica, conociendo el área de una de las 2 mitades, automáticamente se sabe el área de su contraparte. Por ejemplo, tal como se mencionó antes, el área delimitada por la media de la distribución y una desviación estándar a la derecha es de 0.3413 o 34.13%. Se ilustró esto en la figura 6.3. Dada la simetría, el área debajo de la curva normal por encima del eje horizontal y una desviación estándar a la izquierda de la media es también 34.13% del total; sumando ambas áreas se sabe que el área comprendida (siempre se asume que es por debajo de la curva normal y por encima del eje horizontal) entre menos una desviación estándar y una desviación estándar de la media (en otras palabras, una desviación estándar a la izquierda de la media y una desviación estándar a la derecha) contiene el 68.26 del área total o, en términos de probabilidades, representa 68.26% de los casos.

El tipo más común de tabla de áreas bajo la curva normal es el que aparece en este libro y la cual se reproduce en la tabla 6.1. Esta tabla permite determinar valores de área (probabilidad) para porciones de la

Tabla 6.1 Tabla de áreas bajo la curva normal

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990
3.1	0.4990	0.4991	0.4991	0.4991	0.4992	0.4992	0.4992	0.4992	0.4993	0.4993
3.2	0.4993	0.4993	0.4994	0.4994	0.4994	0.4994	0.4994	0.4995	0.4995	0.4995
3.3	0.4995	0.4995	0.4995	0.4996	0.4996	0.4996	0.4996	0.4996	0.4996	0.4997
3.4	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4998

curva normal que se encuentran entre la media y un determinado valor de z, que representa la cantidad de desviaciones estándar a partir de la media. Nótese que se utiliza el símbolo "z" (la letra z minúscula) para representar las desviaciones estándar de la distribución normal estándar, la que tiene $\mu = 0$ y $\sigma = 1$ y que es la que se usa para estandarizar los valores de otras distribuciones normales con valores distintos a éstos para sus medias y sus desviaciones estándar, de manera que sea posible utilizar esta tabla de áreas para cualquier distribución normal. Se ilustra esto en las secciones siguientes.

Como puede verse, la tabla tiene como encabezado la gráfica de la distribución normal, en la que se oscurece una porción entre la media y cierto valor a su derecha. Esto refleja, precisamente, lo que se mencionaba antes respecto a que la tabla permite determinar valores de probabilidad para porciones de la curva normal que se encuentran entre la media y un determinado valor positivo de z.

La tabla tiene en la esquina superior izquierda a la z y en esa columna tiene valores que van de 0.0 y hasta 3.4 y tiene como encabezados del primer renglón, el renglón donde se encuentra la z a la izquierda, valores que van del 0.00 al 0.09, de manera que el valor del renglón da números hasta décimos y el valor de la columna da valores de centésimos por lo que, por ejemplo, en el cruce del 1.0, con la columna del 0.00 se tiene el valor del área correspondiente al 1.00 que es, precisamente, el 0.3413 del que se habla. Esto mismo, en símbolos:

$$P(0 \leq z \leq 1) = 0.3413$$

Este valor de 0.3413 indica que la superficie delimitada entre la media y una desviación estándar por arriba de ella, por encima del eje horizontal y por debajo de la curva normal cubre 34.13% del área total e, interpretando esta área como probabilidad, se puede decir que si se elige al azar un elemento de una población normal con $\mu = 0$ y $\sigma = 1$, la probabilidad de que tenga un valor ubicado entre la media y una desviación estándar por encima de ella es de 0.3413, o 34.13 por ciento.

Aquí vale la pena observar que esta manera de interpretar áreas bajo la curva normal en términos de probabilidades corresponde a la interpretación de la probabilidad como frecuencia relativa que se analizó con detalle en el capítulo 4.

Ahora, como la distribución normal es simétrica, sabemos que el área a la izquierda de la media y a una distancia de una desviación estándar también representa 0.3413 del área:

$$P(-1 \leq z \leq 0) = 0.3413$$

Por supuesto, el área total desde menos una desviación estándar (una desviación a la izquierda de la media) y hasta una desviación estándar (una desviación a la derecha de la media) es 0.6826, la suma de las 2 mitades simétricas:

$$P(-1 \leq z \leq 1) = 0.6826$$

Se ilustran estas 2 áreas en la figura 6.4.

Otro ejemplo: en el cruce del renglón 2.3 con la columna 0.05 se tiene el valor 0.4906, lo cual quiere decir que la proporción de área que se encuentre entre la media de 0, una z de 2.35 (es decir, 2.35 desviaciones estándar a la derecha de la media), por encima del eje horizontal de la gráfica y por debajo de la curva normal es 0.4906, o 49.06%. En símbolos:

$$P(0 \leq X \leq 2.35) = 0.4906$$

Por lo que el área a una distancia de 2.35 desviaciones estándar a ambos lados de la media es:

$$P(-2.35 \leq z \leq 2.35) = 0.992$$

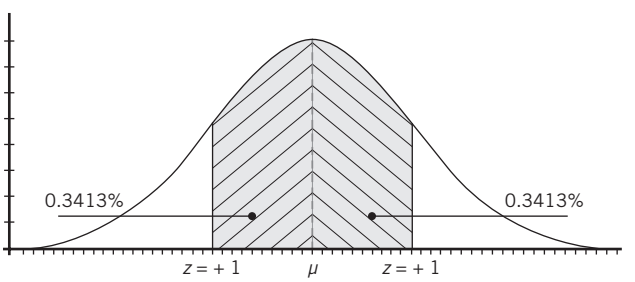


Figura 6.4 El área entre $z = -1$ y $z = 1$ en una distribución normal estándar.

Otros valores que se suelen utilizar para determinar probabilidades con la distribución normal, son a 2 y a 3 desviaciones estándar de la media:

$$P(0 \leq X \leq 2) = 0.4772$$
$$P(-2 \leq X \leq 2) = 0.9544$$
$$P(0 \leq X \leq 3) = 0.4987$$
$$P(-3 \leq X \leq 3) = 0.9974$$

Se ilustran los 3 casos de $z = 1$, $z = 2$ y $z = 3$ en la figura 6.5.

Ahora, como la tabla de áreas bajo la curva normal sólo muestra valores para una de sus 2 mitades simétricas, es necesario comprender cabalmente su uso para determinar diferentes porciones de áreas. Los valores que arroja la tabla son los correspondientes a un área entre la media y un valor positivo de z (el 0.3413) que ya se vio y, por simetría, el valor entre la media y un valor negativo de z (el mismo 0.3413), que mide el área a la izquierda de la media. Se revisan en seguida otros posibles casos de áreas.

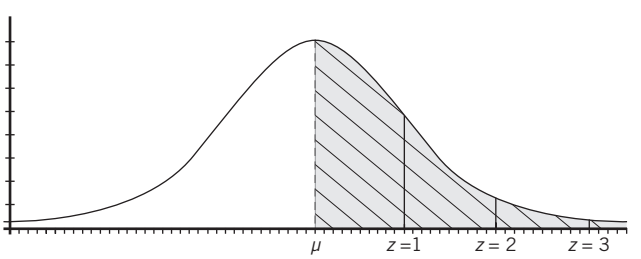


Figura 6.5 Distribuciones normales que muestran las áreas delimitadas por encima del eje horizontal, por debajo de la curva normal y entre *a*) una desviación estándar, *b*) 2 desviaciones estándar y *c*) 3 desviaciones estándar alrededor de la media.

El área en un extremo de la curva normal

Sabiendo que el área entre la media y una desviación estándar a la derecha de la media representa el 0.3413 del área total y sabiendo también que, como se trata de una distribución simétrica, el área total del lado derecho de la media representa la mitad del total, o 0.5, se puede saber fácilmente que el área que se encuentra en el extremo derecho de la gráfica, a partir de una desviación estándar, es igual a 0.5 – 0.3413 = 0.1587. Por supuesto, se aplica la misma idea al área que se encuentra a la izquierda de un valor negativo de z. Esto, en símbolos:

$$P(z \geq 1) = 0.5 - P(0 \leq z \leq 1) = 0.5 - 0.3413 = 0.1587$$
$$P(z \geq -1) = 0.5 - P(0 \leq z \leq -1) = 0.5 - 0.3413 = 0.1587$$

Se ilustran estas áreas en la figura 6.6.

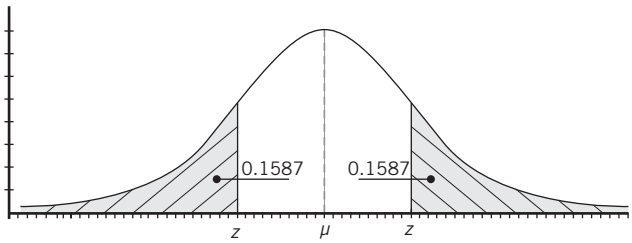


Figura 6.6 Áreas bajo la curva normal, a la derecha de un valor positivo de z y a la izquierda de un valor negativo de z.

El área entre 2 valores positivos de z

También puede ser necesario determinar la probabilidad de que un elemento de una población distribuida normalmente tenga un valor dado entre 2 valores positivos de z. Por ejemplo, el porcentaje de área que se encuentra entre $z = 1.58$ y $z = 2.33$ se encuentra determinando, en primer lugar, el área correspondiente a la porción más grande, el área entre la media y $z = 2.3$, y luego, restándole a ésta la porción más pequeña, el área entre la media y $z = 1.5$.

Así, se encuentra en la tabla de áreas que:

$$P(0 \leq z \leq 2.33) = 0.4901$$
$$P(0 \leq z \leq 1.58) = 0.4429$$

Por lo que: $P(1.58 \leq z \leq 2.33) = 0.4901 - 0.4429 = 0.0472$

Se ilustra este caso en la figura 6.7.

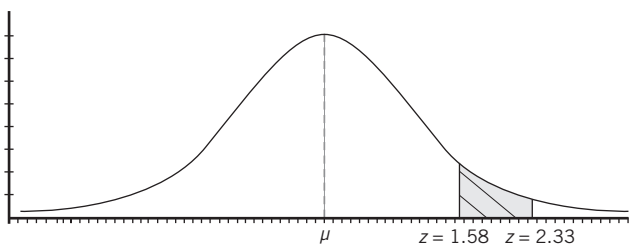


Figura 6.7 El área bajo la curva normal entre 2 valores positivos de z.

El área entre 2 valores negativos de z

Para determinar la probabilidad de que un elemento de una población distribuida normalmente tenga un valor dado entre 2 valores negativos de z se sigue el mismo procedimiento anterior. Por ejemplo, el porcentaje de área que se encuentra entre $z = -2.85$ y $z = -0.66$ se encuentra determinando en primer lugar el área correspondiente a la porción más grande, el área entre la media y $z = -2.85$, y luego, restándole a ésta la porción más pequeña, el área entre la media y $z = -0.66$.

De la tabla de áreas:

$$P(-2.85 \leq z \leq 0) = 0.4978$$
$$P(-0.66 \leq z \leq 0) = 0.2454$$

Por lo que: $P(-2.85 \leq z \leq -0.66) = 0.4978 - 0.2454 = 0.2524$

Se ilustra este caso en la figura 6.8.

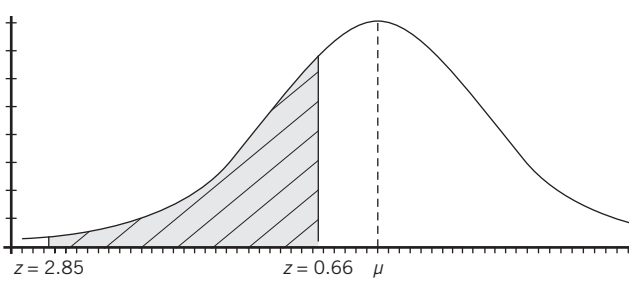


Figura 6.8. El área bajo la curva normal entre 2 valores negativos de z.

El área desde algún punto en la parte izquierda de la media y hasta el extremo derecho de la distribución normal

Incluye toda la mitad derecha de la gráfica y una porción del lado izquierdo de la media por lo que, para calcular la proporción del área total que representa, se deben sumar estas 2 partes. Por ejemplo, para determinar la porción de área que encuentra a la derecha de $z = -1.55$, en primer lugar se determina la porción que se encuentre entre $z = -1.55$ y la media. De la tabla de áreas:

$$P(-1.55 \leq z \leq 0) = 0.4394$$

y como la mitad derecha de la distribución normal representa el 0.5:

$$P(-1.55 \leq z) = P(z \geq -1.55) = 0.5 + 0.4394 = 0.9394$$

Se ilustra este caso en la figura 6.9.

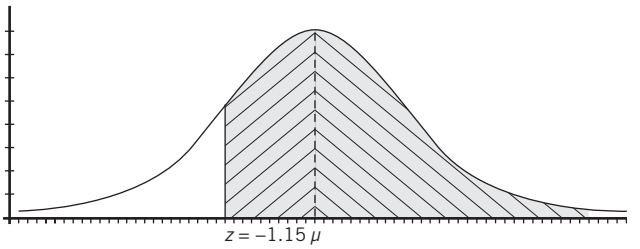


Figura 6.9 El área bajo la curva desde algún punto en la parte izquierda de la media y hasta el extremo derecho de la distribución normal.

El área desde algún punto en la parte derecha de la media y hasta el extremo izquierdo de la distribución normal

Incluye toda la mitad izquierda de la gráfica y una porción del lado derecho de la media, por lo que para calcular la proporción del área total que representa, se deben sumar estas 2 partes. Por ejemplo, para determinar la porción de área que encuentra a la izquierda de $z = 2$, en primer lugar se determina la porción que se encuentre entre $z = 2$ y la media. De la tabla de áreas:

$$P(0 \leq z \leq 2) = 0.4772$$

e incluyendo toda la mitad izquierda del área bajo la curva se tendría:

$$P(-\infty \leq z \leq 2) = 0.9772$$

6.2.3.1 Excel y áreas bajo la curva normal

Excel tiene 4 funciones relacionadas con la distribución normal: 2 funciones que determinan valores para la distribución normal estándar, la que tiene $\sigma = 1$ y $\mu = 0$:

1) DISTR.NORM.ESTAND(z), y
2) DISTR.NORM.ESTAND.INV(Probabilidad)

Además tiene otras 2 que permiten determinar valores para una distribución normal cualquiera, con cualquier media y cualquier desviación estándar:

3) DISTR.NORM(x,media,desv_estándar,acum), y
4) DISTR.NORM.INV(Probabilidad,media,desv_estándar).

Analizaremos estas funciones en seguida:

1) DISTR.NORM.ESTAND(z). Determina la proporción de área que existe entre el extremo izquierdo ($-\infty$) y hasta el valor especificado de z, y es la que se puede utilizar para determinar los valores de las áreas que se determinaron antes con las tablas. En los ejemplos anteriores se vio que el área comprendida entre la media y una desviación estándar (z) a la derecha o la izquierda de la media es 0.3413, resultado que se presentó como:

$$P(0 \leq z \leq 1) = 0.3413$$

Si se utiliza la función de Excel, Distr.Norm.Estand(1) se obtiene 0.84134475, la cual señala que

$$P(-\infty \leq z \leq 1) = 0.84134475$$

La cual es toda el área, desde el extremo izquierdo, incluyendo toda la primera mitad del área y hasta una desviación estándar a la derecha de la media: 0.5 + 0.3413. Se ilustra esto en la figura 6.10.

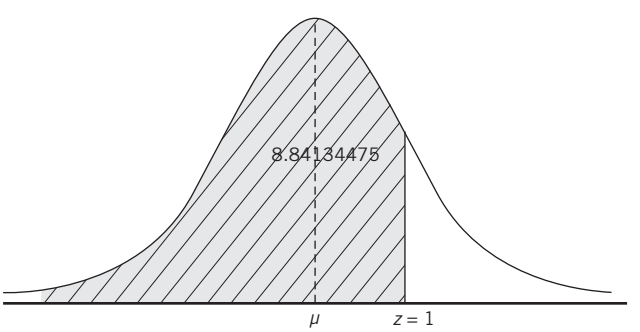


Figura 6.10. El área bajo la curva normal para $P(-\infty \leq z \leq 1) = 0.84134475$.

Con Distr.Norm.Estand(−1) se obtiene 0.15865525, lo cual señala que

$$P(-\infty \leq z \leq -1) = 0.15865525$$

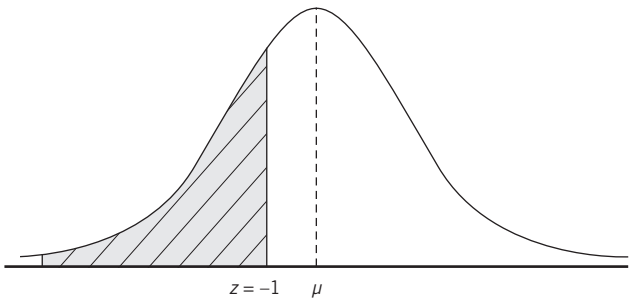


Figura 6.11 El área bajo la curva normal para $P(-\infty \leq z \leq -1) = 0.15865525$.

y es el área desde el extremo izquierdo y hasta una desviación estándar a la izquierda de la media, como se ilustra en la figura 6.11.

Cabe señalar que no existe una función de Excel que determine las probabilidades tal como se presentan aquí en la tabla de áreas, la cual sólo contempla la mitad; pero conociendo la distribución, a partir de este resultado se sabe que de la media a una desviación estándar a la derecha o la izquierda está el 0.34134475 del total del área debajo de la campana.

Con el mismo resultado anterior, es fácil ver que

$$P(z \geq -1) = 1 - 0.15865525 = 0.84134475$$

Lo cual ilustra el caso de un valor del área a la derecha de un valor negativo de z, pero que es, al mismo tiempo, el caso para el área a la derecha de cualquier valor de z positivo o negativo, ya que, como ya se señaló, la función de Excel cubre toda el área de la curva desde menos infinito y hasta el valor de z señalado.

Los ejemplos anteriores muestran el uso de la tabla de áreas bajo la curva normal y los de la función Distr.Norm.Estand de Excel e ilustran 2 maneras de leer las áreas bajo la curva; y existe cuando menos otra: tablas o lecturas que determinan áreas desde el extremo derecho de la curva y no desde el izquierdo como lo hace Excel; sin embargo, como todas ellas son equivalentes, se termina aquí la ilustración de esta función de Excel, no sin antes resaltar el hecho de que los resultados que se obtienen con Excel son mucho más precisos que los que se suelen manejar en las tablas de áreas bajo la curva normal como la que se presenta aquí. Con las tablas se obtienen áreas con precisión de diezmilésimas (4 posiciones decimales) en tanto que con Excel se obtienen a cienmillonésimas (8 posiciones decimales). Cuando la precisión es crucial se vuelve indispensable utilizar Excel (o cálculo integral).

En los ejemplos que se presentan en este texto se asume que la precisión a diezmilésimas es aceptable.

2) DISTR.NORM.ESTAND.INV(Probabilidad). Devuelve el valor de z que delimita cierta porción de área o probabilidad para la distribución normal estándar. Se ilustra su uso en la sección siguiente, la 6.2.3.2.
3) DISTR.NORM(x,media,desv_estándar,acum). Determina la proporción de área (probabilidad) que existe entre el extremo izquierdo ($-\infty$) y hasta el valor especificado de x respecto a la media y la desviación estándar especificadas. Se ejemplifica más adelante el uso de esta función en la sección 6.2.4.1.
4) DISTR.NORM.INV(Probabilidad,media,desv_estándar). Determina el valor de z que divide el área desde $-\infty$ (el extremo izquierdo) de la curva y hasta ese valor de z, de tal manera que el área representa la probabilidad anotada en la función y para la media y la desviación estándar especificadas. Se muestra el uso de esta función en la sección 6.2.4.2.

6.2.3.2 Determinación de *z* a partir del área o la probabilidad

En los ejemplos anteriores se ilustró la manera en la que se puede determinar el porcentaje del área que se encuentra bajo una curva normal (y por encima del eje horizontal) y entre las líneas verticales trazadas sobre 2 puntos cualesquiera y se vio que esta área representa también probabilidad. Para hacer esto, se buscaba en los encabezados de renglón y de columna de la tabla hasta encontrar el valor de z buscado y se encontraba el área correspondiente en el cruce del renglón con la columna.

En muchas aplicaciones resulta también necesario determinar el valor o valores de z que delimita o delimitan cierta porción del área o cierta probabilidad y, en estos casos, se procede a la inversa: se busca en el cuerpo de la tabla el valor de área o de probabilidad que interesa (o el más cercano si no lo hay exacto) y se determina el valor de z identificando los valores correspondientes en el encabezado del renglón y el encabezado de la columna correspondiente. Se revisan en seguida algunos ejemplos.

Se ilustraron antes en la figura 6.5 las áreas entre la media aritmética y los valores de z de 1, 2 y 3, que son valores que se suelen utilizar al manejar esta distribución normal. En el ejemplo siguiente se revisa el caso inverso: la manera en la que determina el valor de z que delimita cierta área a la derecha de la media.

EJEMPLO 6.1

El valor de *z* para un área a la derecha de la media

¿Qué valor de z delimita 30% del área bajo la curva y a la derecha de la media?

Solución:

Se busca el valor de z que delimita el área que se ilustra en la figura 6.12.

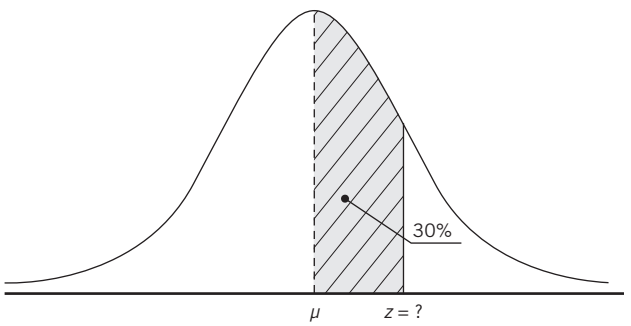


Figura 6.12 Determinación de z para 30% de área a la derecha de la media.

Buscando en el cuerpo de la tabla de áreas bajo la normal, se ve que el valor más cercano a 0.30 es 0.2995 y se encuentra en el renglón de z de 0.8 y en el cruce con la columna de 0.04, por lo que el valor de z correspondiente es 0.84. En la simbología de teoría de la probabilidad: $P(0 \leq z \leq 0.84) = 0.2995$.

Al igual que en este caso, en muchas ocasiones, cuando se desea encontrar el valor de z a partir del área, no se encuentra el valor exacto de ésta, lo cual conduce a 3 opciones:

1. Utilizar el valor más cercano, como se hace en el ejemplo anterior.
2. Interpolar entre los 2 valores que contienen el que se busca, procedimiento que se ilustra en el ejemplo siguiente.
3. Utilizar una función de Excel para determinar el valor correspondiente, ésta puede ser en muchos casos la mejor opción ya que da el resultado exacto, lo cual no sucede con la 1; y es mucho más sencillo que la 2, lo cual quedará ilustrado en el ejemplo siguiente.

EJEMPLO 6.2

Ahora se resuelve el mismo ejemplo anterior, ¿qué valor de z delimita 30% del área bajo la curva y a la derecha de la media? Utilice la interpolación para aproximar más precisamente el resultado de la tabla.

Solución:

Buscando en el cuerpo de la tabla de áreas bajo la normal se ve que no aparece el 0.3000 y que los valores entre los que se encuentra éste son el 0.2995, que corresponde a una z de 0.84 y 0.3023 para una z de 0.85. Para revisar el procedimiento de interpolación se sugiere trazar en primer lugar un diagrama como el de la figura 6.13.

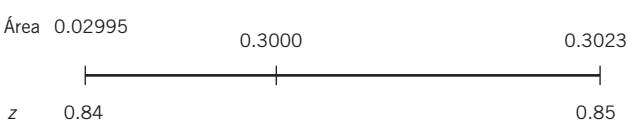


Figura 6.13 Esquema con datos para interpolación.

Se ve en esta figura que el valor preciso de z que se busca está entre 0.84 y 0.85 y se tiene un recorrido de 0.3023 – 0.2995 = 0.0028 entre los 2 valores del área que aparecen en la tabla, y que el 0.30 se encuentra a 0.3000 – 0.2995 = 0.0005 unidades a partir de 0.2995 y estas 5 diezmilésimas representan el 0.0005/0.0028 = 0.178571428, o aproximadamente 17.857% del recorrido total y, como el recorrido de la z va de 0.84 a 0.85, es decir 0.01 (0.85 – 0.84), 17.8571428% de esta centésima es: 0.178571428 (0.01) = 0.00178571428 que, sumado al valor de 0.84 de z da un valor de 0.84178571428. En símbolos:

$$P(0 \leq z \leq 0.84178571428) = 0.30$$

Sin embargo, este valor encontrado mediante interpolación es sólo aproximado, ya que se trata de una aproximación lineal (es decir, en línea recta, como se ilustra en la figura 6.11). El valor exacto se puede determinar mediante cálculo integral que, como ya se mencionó, rebasa el alcance de este libro. La opción que da el valor exacto es la función de Excel DISTR.NORM.ESTAND.INV(probabilidad), que en este caso sería: =DISTR.NORM.ESTAND.INV(0.8), la cual produce como resultado $z = 0.84162123$ y que, como puede verse, difiere del que se encontró mediante interpolación pero que es, no obstante, muy similar.

Comparando los resultados, mediante interpolación se obtuvo:

$$P(0 \leq z \leq 0.84178571428) = 0.30$$

Y, mediante la función de Excel, =DISTR.NORM.ESTAND.INV(0.8), se obtuvo:

$$P(0 \leq z \leq 0.84162123) = 0.30$$

El resultado más preciso es el de Excel pero, aunque el obtenido mediante interpolación es menos preciso que éste, sí es más preciso en comparación con el que se obtiene simplemente utilizando el valor más cercano posible en la tabla, pero, adicionalmente, es posible que esta aproximación con la tabla sea lo suficientemente buena para propósitos prácticos.

En los ejemplos siguientes se ilustran los procedimientos para determinar la z que delimita diversas áreas (probabilidades) de la curva normal. Se ilustró antes en la figura 6.6 el área de 0.1587 en ambos extremos de la curva normal y se encontró que el valor de z correspondiente era $z = \pm 1$. En el ejemplo siguiente se determina el valor de z que delimita otra área en los extremos de la distribución.

EJEMPLO 6.3

El valor de z para un área en un extremo de la curva normal

¿Cuál es el valor de z que delimita 5% del área de la curva normal en sus extremos?

Solución:

En la tabla de áreas bajo la curva normal se encuentra que $P(0 \leq z \leq 1.64) = 0.4495$ y $P(0 \leq z \leq 1.65) = 0.4505$, y como ambos puntos son equidistantes de 0.45 el valor interpolado es simplemente su promedio: el valor aproximado de z mediante interpolación es, entonces, $z = 1.645$. Con Excel: =DISTR.NORM.ESTAND.INV(0.95) = 1.644854, que es el valor exacto; entonces, para ambos extremos:

$$P(0 \leq z \leq 1.644854) = 0.45 \text{ y}$$
$$P(-\infty \leq z \leq -1.644854) = 0.45$$

Se ilustra lo anterior en la figura 6.14.

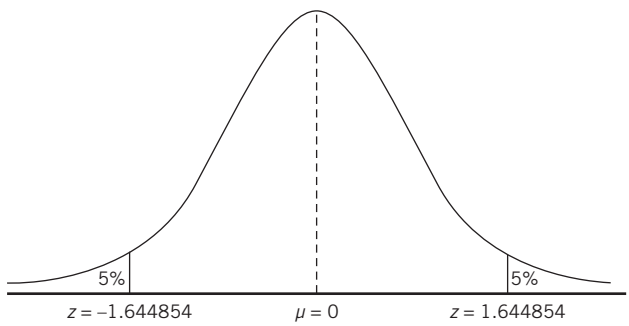


Figura 6.14 Los valores de z que delimitan 5% del área en los extremos de la curva normal.

EJEMPLO 6.4

El valor de z para un área entre 2 valores positivos de z

¿Qué valores de z delimitan 20% del área bajo la curva en medio de la sección derecha de la curva normal?

Solución:

Para resolver este caso, conviene visualizar en una figura la situación. Se ilustra en la figura 6.15. Como puede verse ahí, si se requiere que la porción de 20% esté en medio del lado derecho de la curva, y ésta es la mitad de la campana, entonces debe haber 15% de área entre la media y el inicio de esa porción central y otro 15% a la derecha de la curva.

En la tabla de áreas bajo la curva se encuentra 0.1517 en la z de 0.39 y con Excel se encuentra =DISTR.NORM.ESTAND.INV(0.65) = 0.38532, que es el valor preciso.

Por otra parte, para el valor de z que aísla el 15% del área en el extremo derecho, en la tabla se encuentra que el valor más cercano es $z = 1.03$ para un área de 0.3485 entre la media y ese valor de z, en tanto que con =DISTR.NORM.ESTAND.INV(0.85) se obtiene el valor exacto de $z = 1.036433$.

De lo anterior, entonces, la respuesta a la pregunta de este ejercicio 6.2 se puede resumir como:

$$P(0.38532 \leq z \leq 1.036433) = 0.20$$

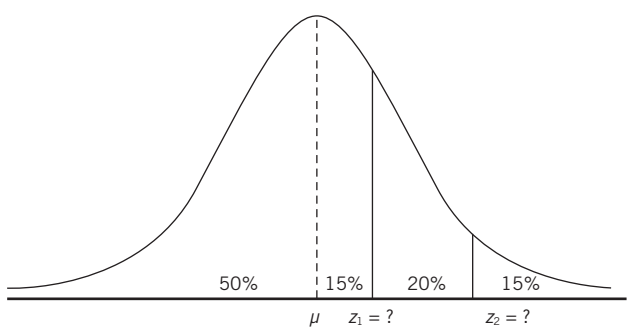


Figura 6.15 Los valores de z que delimitan 20% del área bajo la curva normal, en medio del lado derecho de la curva.

EJEMPLO 6.5

El valor de *z* para un área entre 2 valores negativos de *z*

Por supuesto, se puede plantear la misma pregunta que en el ejercicio anterior, pero ya no para el extremo derecho, sino el izquierdo: ¿qué valores de z delimitan 20% del área bajo la curva en medio de la sección izquierda de la curva normal?

Solución:

Tras resolver el ejemplo anterior, la respuesta a esta pregunta es, aprovechando la simetría de la curva, sencillamente la contraparte de la anterior:

$$P(-1.036433 \leq z \leq -0.38532) = 0.20$$

Se ilustra esto en la figura 6.16.

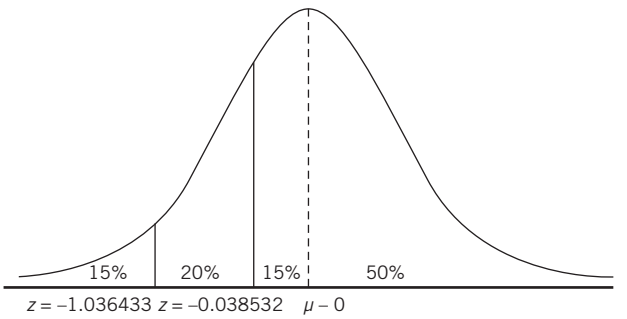


Figura 6.16 Valores de z que delimitan 20% del área bajo la curva, a la izquierda de la media.

EJEMPLO 6.6

El valor de *z* que delimita un área desde la parte izquierda de la media y hasta el extremo derecho de la distribución normal

¿Qué valor de z delimita 60% del área en el extremo derecho de la curva normal?

Solución:

Este valor de z es negativo ya que el área correspondiente incluye toda la mitad derecha de la curva y a partir de 10% a la izquierda de la media. De las tablas de áreas bajo la curva normal, el valor más cercano de z para un área de 0.10 es $z = 0.25$ y corresponde a un área de 0.0987 o, en símbolos:

$$P(-0.250 \leq z \leq \infty) = 0.5987$$

De manera exacta con Excel, si se quiere 60% del extremo derecho y la función de Excel calcula desde el extremo izquierdo, se tiene que: DISTR.NORM.ESTAND.INV(0.40) = –0.25335. En símbolos:

$$P(-0.25335 \leq z \leq \infty) = 0.60$$

EJERCICIOS 6.2.3 Tabla de áreas bajo la curva normal

Determinación del área para determinados valores de *z*

1. Determine el área que se encuentra entre $z = -1.75$ y $z = 2$
2. Indique el área que se encuentra entre $z = -2.3$ y $z = -0.82$
3. Precise el área que se encuentra en $z \geq -1.25$
4. Defina el área que se encuentra en $z \geq 0.97$
5. Señale el área que se encuentra en $z \leq 2.74$
6. Determine el área que se encuentra en $z \leq -2.33$ y $z \geq 2.33$
7. Indique el área que se encuentra entre $z = 1.97$ y $z = 2.05$
8. Precise el área que se encuentra entre $z = -0.03$ y $z = 1.17$
9. Defina el área que se encuentra en $z \leq -2.27$
10. Señale el área que se encuentra entre $z = -1.43$ y $z = -0.93$

Determinación de *z* a partir del área o la probabilidad

11. ¿Qué valor de z delimita 40% del área bajo la curva y a la derecha de la media?
12. ¿Qué valor de z delimita 10% del área bajo la curva y a la izquierda de la media?
13. ¿Cuál es el valor de z que delimita 25% del área de la curva normal en sus extremos?
14. ¿Cuál es el valor de z que delimita 50% del área de la curva normal en sus extremos?
15. ¿Qué valores de z delimitan 10% del área bajo la curva en medio de la sección derecha de la curva normal?
16. ¿Qué valores de z delimitan 40% del área bajo la curva en medio de la sección derecha de la curva normal?
17. ¿Qué valores de z delimitan 3% del área bajo la curva en medio de la sección izquierda de la curva normal?
18. ¿Qué valores de z delimitan 7.5% del área bajo la curva en medio de la sección izquierda de la curva normal?

19. ¿Qué valor de z delimita 0.5% del área en el extremo derecho de la curva normal?

20. ¿Qué valor de z delimita 15.5% del área en el extremo izquierdo de la curva normal?

6.2.4 Determinación de probabilidades para cualquier distribución normal

En la sección anterior se vio cómo se pueden determinar áreas bajo la curva para cualquier porción de una distribución normal estándar, la que tiene media de 0 y desviación estándar de 1.

Como es prácticamente infinita la cantidad de distribuciones normales que se pueden encontrar o construir combinando diferentes valores de la media y de la desviación estándar, para encontrar probabilidades en estos casos se debe estandarizar la medida, expresando la distancia entre el punto de interés, X, y la media correspondiente en unidades de la desviación estándar, de la siguiente manera:

$$z = \frac{X - \mu}{\sigma} \tag{6.3}$$

Se ilustra el procedimiento en el siguiente ejemplo.

EJEMPLO 6.7

En una envasadora de agua purificada se producen botellas que tienen, en promedio, medio litro de agua, con una desviación estándar de 5 ml (mililitros). Si se sabe por experiencia que la distribución de los contenidos de agua de estas botellas tiene distribución normal, determinar qué porcentaje de las botellas que se fabrican tiene entre 490 y 510 ml.

Solución:

En primer lugar, y como una recomendación general, se sugiere elaborar una gráfica de la distribución normal para comprender cabalmente las circunstancias. Así, en la figura 6.17 se muestra una gráfica en la que se ha colocado al centro la media de 500 ml, a sus lados se marcó sobre el eje horizontal los valores de 490 y 510 desde donde se trazan 2 rayas verticales que delimitan el área de interés, misma que se achuró.

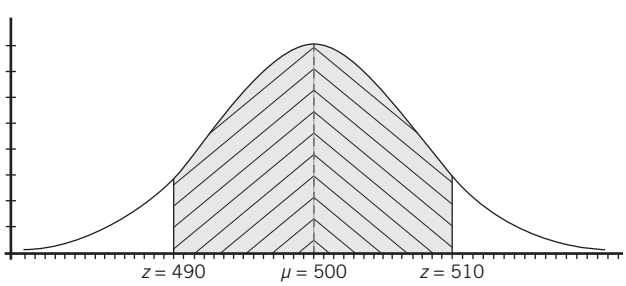


Figura 6.17 Gráfica con los datos del ejemplo 6.7.

Tal como puede apreciarse en la gráfica, interesa el área que está entre 500 y 510 ml y entre 490 y 500 ml y se puede ver, al mismo tiempo y recordando la simetría de la distribución normal, que éstas dos áreas son, a su vez, simétricas por lo que, conociendo la proporción de área que cualquiera de ellas representa, se sabe automáticamente que la contraparte tiene la misma área. Entonces, calculando la z para el área a la derecha de la media:

$$z = \frac{X - \mu}{\sigma} = \frac{510 - 500}{5} = 2$$

Entonces, este valor de $z = 2$ indica que el valor 510 ml está a 2 desviaciones estándar de la media con lo que, consultando la tabla de áreas bajo la curva normal, se encuentra que:

$$P(0 \le z \le 2) = P(500 \le X \le 510) = 0.4772.$$

Por su parte, el área del lado izquierdo de la media:

$$z = \frac{X - \mu}{\sigma} = \frac{490 - 500}{5} = -2$$

Como sabemos por la simetría de la distribución normal, el área del lado izquierdo es igual a la del lado derecho, así que:

$$P(-2 \le z \le 0) = P(490 \le X \le 500) = 0.4772$$

Entonces: $P(-2 \le z \le 2) = P(490 \le X \le 510) = 0.9544$

En el ejemplo 6.8 se revisa el caso cuando se conoce el porcentaje de área y lo que se busca es el valor de la variable que divide el área de la curva normal con respecto a ese porcentaje. Sin embargo, en la sección siguiente se comienza por revisar algunos detalles adicionales de las funciones de Excel para la distribución normal.

6.2.4.1 Excel y probabilidades para cualquier distribución normal

La función DISTR.NORM(x,media,desv_estándar,acum) permite determinar áreas, es decir, probabilidades, para cualquier distribución normal, conociendo su media y su desviación estándar y es la única de las 4 funciones de Excel para la normal que permite elegir si desean probabilidades acumuladas o no. En el primer caso se deberá anotar 1 en el lugar del último parámetro. Si se anota 0, o se omite, Excel produce la probabilidad no acumulada.

Resolviendo el mismo ejemplo 6.7, DISTR.NORM(490,500,5,1) arroja como resultado 0.02275013 que, como ya se vio, representa la proporción de área desde el extremo izquierdo y hasta el valor de 490, para una población con media de 500 y desviación estándar de 5, o sea: $P(-\infty \leq X \leq 490) = 0.02275013$, de donde $P(490 \leq X \leq 500) = 0.5 - 0.02275013 = 0.47724987$, que es casi el mismo valor que se encontró en las tablas de áreas en el ejemplo, sólo que, como ya también se mencionó, más exacto. De donde,

$$P(-2 \leq z \leq 2) = P(490 \leq X \leq 510) = 0.47724987(2) = 0.95449974$$

Nótese que para usar esta función de Excel fue necesario anotar un "1" como el último parámetro (DISTR.NORM(490,500,5,1)) para hacer que el programa devuelva la probabilidad acumulada hasta ese punto.

Si se omite el 1 o se anota 0 en su lugar, lo que se obtiene es la "función de masa de masa de probabilidad" o, en otras palabras, la ordenada de la curva en ese punto. Para ilustrar qué es esto, se tabuló esta función para valores de 485 y hasta 515, de uno en uno y utilizando la media de 500 y desviación estándar de 5, como en el ejemplo. En la tabla 6.2 se presentan los resultados. Y en la figura 6.18 se graficaron estos resultados.

Tabla 6.2 Resultados obtenidos para Distr.Norm (x,500,5,0)

Valor de *x*	Resultado
485	0.00088637
486	0.00158309
487	0.00271659
488	0.00447891
489	0.00709492
490	0.01079819
491	0.01579003
492	0.02218417
493	0.02994549
494	0.03883721
495	0.04839414
496	0.05793831
497	0.06664492
498	0.07365403
499	0.07820854
500	0.07978846

Valor de *x*	Resultado
501	0.07820854
502	0.07365403
503	0.06664492
504	0.05793831
505	0.04839414
506	0.03883721
507	0.02994549
508	0.02218417
509	0.01579003
510	0.01079819
511	0.00709492
512	0.00447891
513	0.00271659
514	0.00158309
515	0.00088637

Como puede verse en la gráfica, la función produce los valores de las ordenadas (los valores *Y*, que se miden sobre el eje vertical) con los que se delinea la curva normal correspondiente a esa media y esa desviación estándar. Por supuesto, si se hace lo mismo utilizando 0 como la media y 1 como la desviación estándar, se obtienen los datos y la gráfica de la distribución normal estándar. Se continúa ahora con los ejemplos de cálculo de valores de la variable de distribuciones normales a partir de los datos de probabilidad o área.

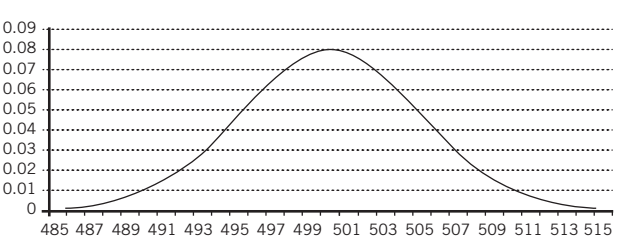


Figura 6.18 Gráfica de la distribución normal mediante la función Distr.Norm(x,500,5,0).

EJEMPLO 6.8

En una población de observaciones que tiene distribución normal y desviación estándar de 10, la probabilidad de que una observación elegida al azar sea mayor de 50 es de 20%. Determine: *a*) la media aritmética de la población y *b*) el valor por encima del cual se encuentra 5% del total de las observaciones.

Solución:

a) En la figura 6.19 se ilustran las condiciones planteadas.

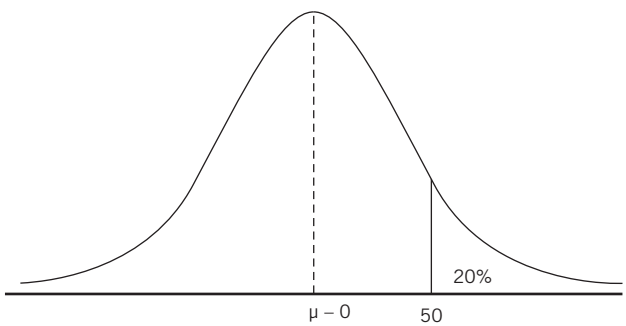


Figura 6.19 Condiciones para el ejemplo 6.8.

Para encontrar la media se parte de la fórmula de z, que la incluye:

$$z = \frac{X - \mu}{\sigma}$$

y, como en este caso conocemos tanto $X = 50$ como $\sigma = 10$, sólo falta determinar el valor de z para poder despejar la media.

Como se sabe que se tiene 20% del área en el extremo derecho de la curva, automáticamente se desprende que el área entre la media y el valor de 50 tiene 30% restante de esta mitad derecha de la curva. Así, se busca el valor 0.3000 en el cuerpo de la tabla de áreas bajo la curva y se encuentra 0.2995, que es el valor más cercano a 0.3000. Este valor se encuentra en el cruce del renglón 0.8 con la columna 0.04, por lo que el valor correspondiente de z es 0.84. En símbolos:

$$P(0 \le z \le 0.84) = 0.2995$$

Con Excel, la función =DISTR.NORM.ESTAND,INV(0.20) da como resultado –0.84162123, en donde el signo negativo indica que es una z a la izquierda de la media.

Conociendo este valor de z se puede ahora despejar en la fórmula:

$$z = \frac{X - \mu}{\sigma}$$

$$0.84 = \frac{50 - \mu}{10}$$

$$50 - \mu = 0.84(10) = 8.4$$

$$-\mu = 8.4 - 50 = -41.6$$

Por lo que la media es igual a 41.6 o, en símbolos, $\mu = 41.6$.

b) Conociendo el valor de la media se puede determinar el valor por encima del cual se encuentra 5% de los valores de esta distribución normal, y de nueva cuenta, como se conoce el porcentaje de área (5%) que divide la porción de interés, se sabe que la parte entre la media y este 5% representa 45% de la mitad derecha del área por lo que se busca el valor 0.4500 de nuevo en el cuerpo de la tabla y se encuentra que el 0.4495 se localiza en el cruce que marca $z = 1.64$ y el 0.4505 está en el cruce que señala $z = 1.64$, por lo que puede utilizarse el promedio de ambos valores como la mejor aproximación del valor de z que divide la mitad del área bajo la curva en 45 y 5%. En símbolos:

$$P(z \ge 1.645) = 0.05$$

Conociendo este valor de z se sustituye en la fórmula de z y se despeja el valor de X:

$$z = \frac{X - \mu}{\sigma}$$

$$1.645 = \frac{X - 41.6}{10}$$

$$16.45 + 41.6 = X$$

$$X = 58.05$$

Resumiendo: $P(X \ge 58.05) = 0.05$

EJEMPLO 6.9

Se sabe que el ciclo de vida de un tipo de bacterias sigue una distribución normal con una media de 200 horas y una desviación estándar de 20 horas. ¿Cuál es la probabilidad de que una bacteria aleatoriamente seleccionada dure entre 200 y 240 horas?

Solución:

En la figura 6.20 se ilustran las condiciones planteadas.

$$z = \frac{X - \mu}{\sigma} = \frac{240 - 200}{20} = 2$$

$$P(0 \le z \le 2) = 0.4772$$

o sea 47.72% y esta probabilidad se expresa en unidades de z, la desviación estándar de la distribución normal estándar. Ahora en las unidades originales:

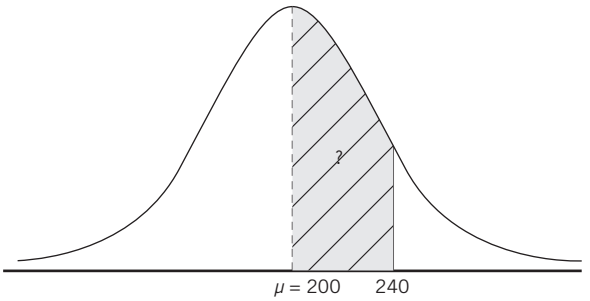


Figura 6.20 Condiciones del ejemplo 6.9.

$$P(200 \leq X \leq 240) = 0.4772$$

Lo cual significa que la probabilidad de que una bacteria aleatoriamente seleccionada dure entre 200 y 240 horas es de 47.72 por ciento.

EJEMPLO 6.10

Las cajas de una marca de cereal contienen en promedio 85 gramos, con una desviación estándar de 2.5 gramos. Si se sabe que el proceso de llenado sigue una distribución normal, ¿cuál es la probabilidad de que una caja contenga entre 82 y 84 gramos?

Solución:

En la figura 6.21 se ilustran las condiciones planteadas.

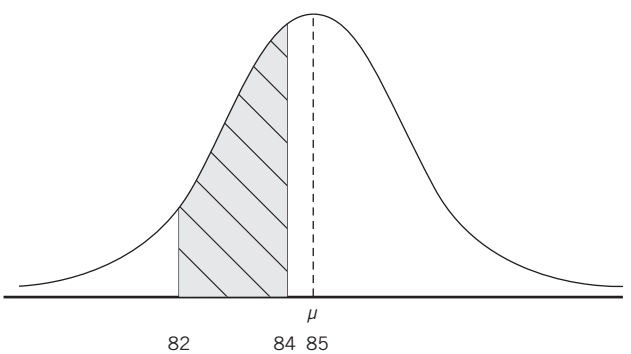


Figura 6.21 Condiciones del ejemplo 6.10.

El área entre la media y 82:

$$z = \frac{X - \mu}{\sigma} = \frac{82 - 85}{2.5} = -1.2$$

$$P(-1.2 \leq z \leq 0) = P(82 \leq X \leq 85) = 0.3849$$

El área entre la media y 84:

$$z = \frac{X - \mu}{\sigma} = \frac{84 - 85}{2.5} = -0.4$$

$$P(-0.4 \leq z \leq 0) = P(84 \leq X \leq 85) = 0.1554$$

Por lo que el área entre 82 y 84:

$$P(-1.2 \leq z \leq -0.4) = P(82 \leq X \leq 84) = 0.3849 - 0.1554 = 0.2295$$

En este resultado se combinaron las expresiones de probabilidad en términos de z y en términos de la variable original, X. Este planteamiento final se puede interpretar diciendo que la probabilidad de que z esté entre -1.2 y -0.4 es de 29.95% y, a su vez, es igual a la probabilidad de que una caja elegida aleatoriamente contenga entre 82 y 84 gramos de cereal.

EJEMPLO 6.11

Para poder ingresar a una escuela preparatoria, un estudiante requiere responder correctamente al menos 75% de las preguntas del examen de admisión, si los resultados siguen una distribución normal, con una media de 63 aciertos y una desviación estándar de 4, ¿cuál es el mínimo de aciertos que debe tener un estudiante para ingresar a la escuela?

Solución:

En la figura 6.22 se ilustran las condiciones planteadas.

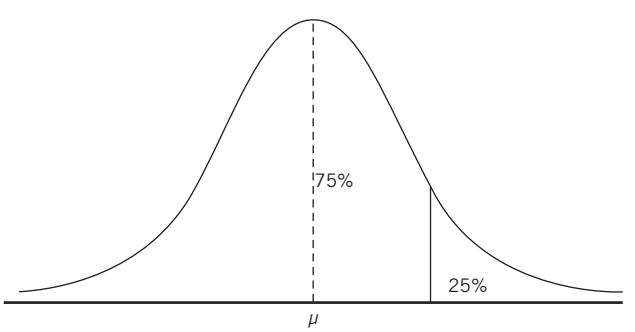


Figura 6.22 Condiciones del ejemplo 6.11.

Con Excel, =DISTR.NORM.ESTAND.INV(0.75), produce 0.67448975, que es el valor de z que divide el área en el 75% en el extremo inferior y el 25% en el superior.

Sustituyendo en la fórmula para estandarizar los valores en las unidades originales:

$$0.674489755 = \frac{X - 63}{4}$$

$$= 0.674489755(4) + 63$$

$$= 65.7$$

El mínimo de aciertos que debe tener un estudiante para ingresar a la escuela es de 66.

Ejemplo 6.12

La vida útil de un modelo de llantas tiene una desviación estándar de 5 000 km y la probabilidad de que dure más de 47 000 kilómetros es de 35 por ciento.

a) Calcule la media.
b) Calcule los kilometrajes alrededor de la media entre los que se encuentra 48% de las observaciones.

Solución:

En las figuras 6.23 y 6.24 se ilustran las condiciones planteadas en el inciso *a*).

a)

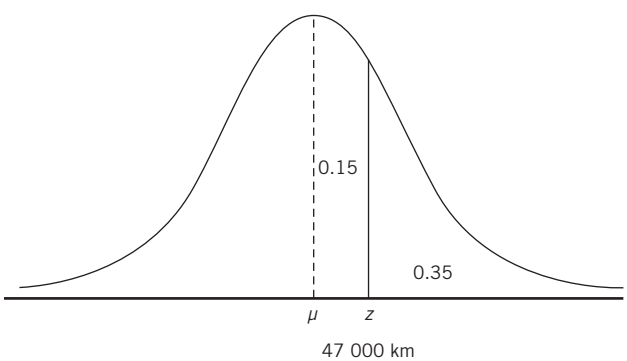


Figura 6.23 Condiciones para *a*) del ejemplo 6.12.

De la tabla de áreas bajo la curva normal,

$$P(0.39 \le z \le 0) = 0.15$$

de donde,

$$0.39 = \frac{47\,000 - \mu}{5\,000}$$

$$47\,000 - \mu = 0.39\,(5\,000)$$

$$\mu = 47\,000 - 1\,950$$

$$\mu = 45\,050$$

b)

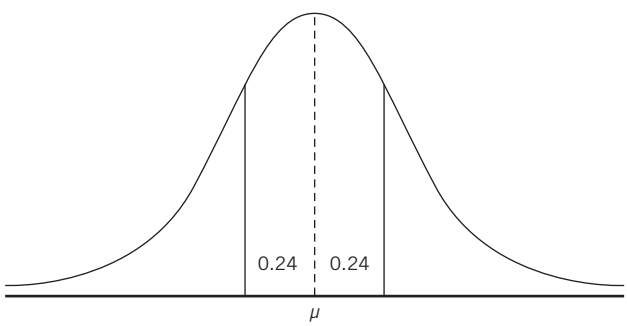


Figura 6.24 Condiciones para *b*) del ejemplo 6.12.

De la tabla de áreas bajo la curva normal,

$$P(0 \le z \le 0.64) = 0.24$$

de donde,

$$0.64 = \frac{X - 45\,050}{5\,000}$$

$$X = 0.64(5\,000) + 45\,050 = 3\,200 + 45\,050$$

$$X = 48\,250$$

Dada la simetría de la curva, $X_1 = 45\,050 - 3\,200 = 41\,850$, con lo que:

$$P(41\,850 \le X \le 48\,250) = 0.48$$

Lo cual quiere decir, en otros términos, que 48% de las llantas tienen una duración de entre 41 850 y 48 250 kilómetros.

6.2.4.2 Determinación de valores de la variable, o de *z*, a partir del área o la probabilidad

En los ejemplos anteriores se ilustró la forma en que se determinan probabilidades para poblaciones con distribución normal, a partir de su media y su correspondiente desviación estándar. En esta sección se

presentan ejemplos de cómo se encuentran valores de una variable con distribución normal a partir de datos sobre probabilidades las cuales, como ya se ha ilustrado, también corresponden a áreas bajo la curva normal.

Ejemplo 6.13

La renta mensual promedio de las casas habitación disponibles en arrendamiento en una zona de la ciudad es de $3 500, con una desviación de $540. Calcule el valor debajo del cual se encuentra el 10% que paga menos.

Solución:

En la figura 6.25 se ilustran las condiciones.

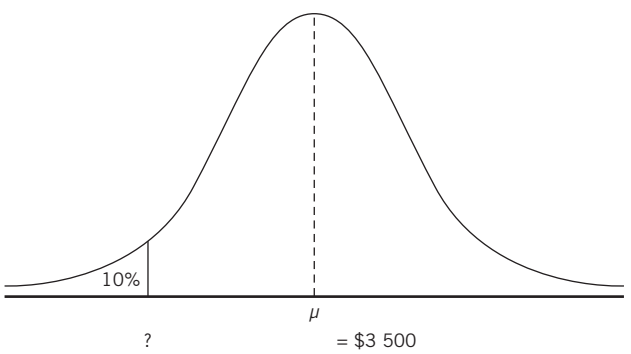


Figura 6.25 Condiciones para el ejemplo 6.13.

En la figura anterior se puede ver que ese 10% menor se puede medir en términos de z y en términos de la variable (renta mensual). El valor correspondiente de z se obtiene de la tabla de áreas bajo la curva normal:

$$P(z \leq -1.285) = 0.1$$

Con Excel, =DISTR.NORM.ESTAND.INV(0.10) se obtiene el mismo valor de $z = -1.28155$.

A partir de la relación de z con los valores de la variable y su media y su desviación estándar, se puede plantear que:

$$-1.285 = \frac{X - 3\,500}{540}$$

de donde, despejando X, se obtiene,

$$X - 3\,500 = -1.285(540) = 692.037$$
$$X - 692.037 + 3\,500 = \$2\,807.963$$
$$P(X \leq 2\,807.96) = 0.1$$

Lo cual quiere decir que el pago máximo de renta de los arrendatarios que menos pagan por su casa habitación en esa zona de la ciudad es de $2 807.96.

Ejemplo 6.14

El consumo promedio de cerveza por año y por persona en cierta población tiene una desviación estándar de 17 litros. Si la probabilidad de que una persona consuma más de 60 litros es de 12%, determine: *a*) la media aritmética de la población y *b*) el número mínimo de litros de cerveza que bebe 5% de los bebedores que más beben cerveza.

Solución:

En la figura 6.26 se representan las condiciones de este ejemplo.

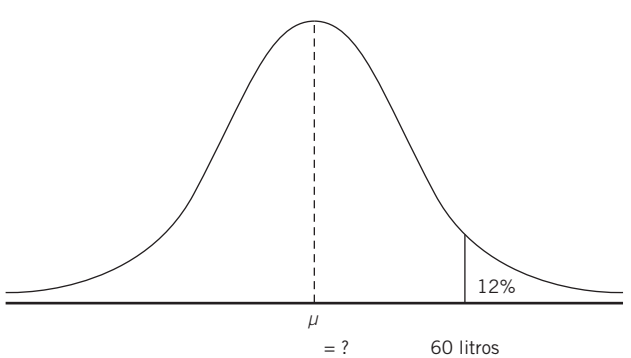


Figura 6.26 Condiciones del ejemplo 6.14.

De la tabla de áreas bajo la curva normal, sabemos que

$$P(0 \leq z \leq 1.17) = 0.38$$

Con Excel, la función =DISTR.NORM.ESTAND.INV(0.88) da como resultado $z = 1.174987$.

O sea que 1.17 es el valor de z que aísla 12% del área a la derecha de la curva, que es el área de la curva que representa a los miembros de esa población que beben más de 60 litros de cerveza al año.

a) Ahora, conociendo estos datos, y a partir de la relación de z con los valores de la variable y su media y su desviación estándar, se puede plantear:

$$1.17 = \frac{60 - \mu}{17}$$

Despejando la media se encuentra que

$$60 - \mu = 1.17(17) = 19.89$$
$$-\mu = 19.89 - 60 = -40.11$$

o sea que $\mu = 40.11$.

b) Ahora, para el mínimo de litros de cerveza que bebe 5% de los bebedores que más bebe, de la tabla de áreas bajo la curva normal se desprende que

$$P(z \geq 1.645) = 0.05$$

Con Excel, la función =DISTR.NORM.ESTAND.INV(0.95) da como resultado $z = 1.644854$.

De la relación de z con los valores de la variable y su media y su desviación estándar, se tiene que:

$$1.645 = \frac{X - 40.11}{17}$$

$$27.965 + 40.11 = X$$

$$X = 68.075$$

En símbolos:

$$P(X \geq 68.075) = 0.05.$$

ejercicios 6.2.4 Determinación de probabilidades para cualquier distribución normal

Determinación de probabilidades o áreas

1. Si la distribución de los salarios semanales de 100 trabajadores es normal y tiene una media de $1 100 y una desviación estándar de $65; si se elige al azar a un trabajador, ¿cuál es la probabilidad de que:

a) tenga un salario igual o inferior a $1 005?
b) su salario sea entre $1 000 y $1 200?
c) tenga un salario de $3 000 o más?

2. El mexicano adulto tiene una estatura promedio de 167 centímetros, con una desviación estándar de 3 centímetros. Si se elige a un adulto mexicano al azar:

a) ¿Cuál es la probabilidad de que mida menos de 160 centímetros?
b) ¿Cuál es la probabilidad de que mida más de 170 centímetros?

3. Según una encuesta, los trabajadores de oficina de una empresa pasan 28 horas por semana trabajando en una computadora. Si la desviación estándar es de 7 horas y se sabe que sigue una distribución normal:

a) ¿Cuál es la probabilidad de que un trabajador elegido al azar utilice la computadora menos de 10 horas?
b) ¿Cuál es la probabilidad de que la utilice entre 40 y 45 horas?

4. El tiempo promedio para leer un periódico es de 48 minutos, la desviación estándar es de 16 minutos. Si los tiempos de lectura tienen distribución normal:

a) ¿Cuál es la probabilidad de que una persona tarde al menos 65 minutos en leer el periódico?
b) ¿Cuál es la probabilidad de que no tarde más de 35 minutos?
c) ¿Cuánto tiempo les toma a 10% de las personas que más se tardan en leer el periódico?

5. El tiempo necesario para realizar una prueba de aptitud es de 85 minutos con una desviación estándar de 15 minutos.

a) ¿Cuál es la probabilidad de que una persona cualquiera termine en 60 minutos o menos?
b) ¿Cuál es la probabilidad de que termine en 65 y 75 minutos?

6. El departamento de recursos humanos de una inmobiliaria calcula que los aspirantes al puesto de ejecutivo de cuenta que hacen examen de ingreso obtienen, en promedio, 84 puntos, con una desviación estándar de 5. Los resultados señalan que las calificaciones siguen una distribución normal. Determine qué porcentaje de aspirantes obtuvieron entre 75 y 95 puntos.

7. El departamento de finanzas estima que los miércoles, día en que se paga a los proveedores, se utilizan $3 200 000.00 con una desviación estándar de $80 000.00. Si el departamento quiere evitar recursos ociosos en la cuenta de cheques y si la distribución de estos pagos es normal, ¿qué probabilidad hay de que en un miércoles cualquiera se paguen más de $3 300 000.00?

8. Un tesorero se dio cuenta de que, en promedio, el saldo semanal en bancos es de $345 000.00, con una desviación estándar de $10 500.00. Si sabe por experiencia que los saldos tienen una distribución normal, determine qué porcentaje de semanas tienen un saldo menor a $320 000.00.

9. Un artículo de una revista de negocios indica que las 21 empresas más importantes de determinado país repartieron este año, en promedio, $25 millones como dividendos en efectivo. Si se toma en cuenta que esta variable tiene distribución normal con desviación estándar de 0.9 millones, determine qué porcentaje de las empresas repartieron entre 25.5 y 26.5 millones.

10. Un asesor fiscal gana en promedio al mes $32 000 con una desviación estándar de $2 500. ¿Qué probabilidad hay de que el siguiente mes gane menos de $32 800?

Determinación de *z* a partir del área o la probabilidad

11. En cierta población, las personas ven la televisión en promedio 4.3 horas al día, con una desviación estándar de 1.2 horas. Calcule el número de horas por encima del cual se encuentra 15% que más ve la televisión.

12. La cantidad de litros de gasolina que los automovilistas compran al entrar en una gasolinera tiene una desvia-

ción estándar de 3.4. La probabilidad de que compren menos de 30 litros es de 7%. Determine:

a) La media aritmética de la población.
b) El valor por encima del cual se encuentra 9% del total de las observaciones.

13. El tiempo que tarda en surtir efecto un medicamento tiene una desviación estándar de 4 minutos. La probabilidad de que tarde más de 11 minutos es de 20%. Determine:

a) La media aritmética de la población.
b) El valor por encima del cual se encuentra 25% del total de las observaciones.

14. El número de piezas empaquetadas por un trabajador en cada turno tiene una desviación estándar de 8, si la probabilidad de que empaque más de 35 piezas es de 30%. Calcule:

a) La media.
b) El número de piezas empaquetadas alrededor de la media entre los que se encuentra 46% de las observaciones.

15. Se realizó un examen para medir el nivel de aprovechamiento de los alumnos de 1o. de preparatoria, se considera que un alumno cuenta con los conocimientos básicos para este nivel educativo si responde correctamente 63% del examen. Con una media de 105 aciertos y una desviación estándar de 5:

a) ¿Cuál es el mínimo de aciertos que debe obtener un alumno para considerar que cuenta con los conocimientos básicos?
b) ¿Qué número de aciertos obtuvo 15% con menor rendimiento?

16. En promedio cada mezcladora de una planta productora de detergente para ropa, genera 87 kilos de desechos al día, con una desviación estándar de 11.4.

a) ¿Cuál es la probabilidad de que se produzcan entre 73 y 81 kilos?
b) ¿A partir de cuántos kilos de desechos genera 12% que más desechos genera?

17. En promedio se deforestan 540 m^2 bimestralmente de un bosque a causa de la tala inmoderada, con una desviación estándar de 27.8 m^2. En 18% de periodos en que se presenta la menor deforestación, ¿cuál es el máximo de metros cuadrados que se deforestan?

6.3 Ajuste cuando se utiliza la distribución normal para evaluar probabilidades de una variable discreta (ajuste por discontinuidad)

Como la normal es una distribución continua, cuando se aplica a casos en los que se tiene una variable discontinua, o discreta, se debe hacer un ajuste, como se explica en seguida con el ejemplo.

Ejemplo 6.15

Los expertos de una casa de bolsa estiman que en un día normal se negocian en promedio 1 800 acciones de la empresa ABC en la bolsa de valores, con una desviación estándar de 67 títulos. Si estos expertos determinaron que el flujo de operaciones tiene una distribución normal, determine en qué porcentaje de días se negocian entre 1 800 y 1 934 acciones.

La solución, con el procedimiento que se ilustró antes, sería:

$$z = \frac{X - \mu}{\sigma} = \frac{1\,934 - 1\,800}{67} = 2$$

En la figura 6.27 se ilustra esto.

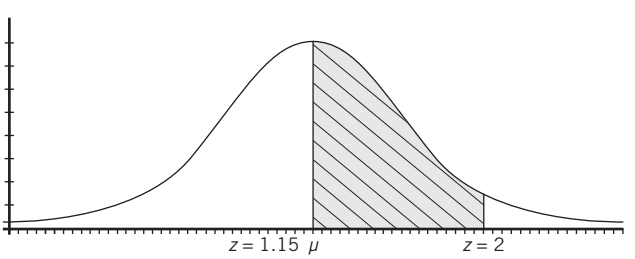


Figura 6.27 Condiciones del ejemplo 6.15.

De la tabla de áreas bajo la curva normal:

$$P(0 \le z \le 2) = 0.4772 = 47.72\%$$
$$P(1\,800 \le X \le 1\,934) = 47.72\%$$

Sin embargo, la variable en este caso, número de acciones negociadas, es discontinua: 1, 2, 3,..., acciones negociadas, por lo que la aplicación de ese procedimiento no es enteramente precisa.

Si los valores posibles de la variable son sólo números enteros, entonces se puede pensar que el 1 800 abarca en realidad desde el 1 799.5, que es donde termina el número 1 799, y hasta el 1 800.5, que es donde empieza el 1 801. Se ilustra esto en la figura 6.28.

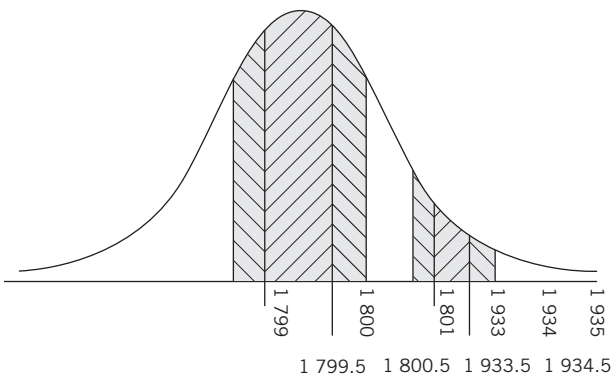


Figura 6.28 Ilustración de la discontinuidad en el ejemplo 6.15.

Como puede verse en la figura, al tratarse de una variable discontinua, cada valor conforma una barra completa en la gráfica, de manera que lo que se conoce como *límites reales* entre un valor y otro se encuentra en medio; por ejemplo, los límites reales de 1 800 son 1 799.5 y 1 800.5; por ello, cuando se determinan probabilidades de variables discontinuas utilizando la distribución normal, que es continua, se deben tomar en cuenta estos límites reales y, entonces, los cálculos adecuados son:

$$z = \frac{X - \mu}{\sigma} = \frac{1\,934.5 - 1\,799.5}{67} = 2.0149$$

y se tomó el valor 1 799.5 porque es el extremo izquierdo del área de interés, al igual que el 1 934.5 es su extremo derecho. Con este valor para z de la tabla de áreas bajo la curva normal se tiene que:

$$P(0 \le z \le 2.01) = 0.4778 = 47.78\%$$
$$P(1\,800 \le X \le 1\,934) = 47.78\%$$

Estos resultados indican que, sin el ajuste por discontinuidad, la probabilidad de que en un día elegido al azar el número de acciones negociadas esté entre 1 800 y 1 934 es de 47.72%, en tanto que, utilizando el ajuste, la probabilidad se convierte en 47.78 por ciento.

Aunque esta diferencia pueda parecer minúscula, los resultados en términos de las decisiones que se tomen con base en esta información pueden ser considerables; por ello, se recomienda siempre utilizar este ajuste cuando se trabaja la normal con variables discontinuas.

EJEMPLO 6.16

Una empresa de seguridad privada recibe de sus clientes 15 quejas a la semana en promedio, con una desviación estándar de 3. ¿Qué probabilidad hay de que esta semana se reciban entre 10 y 14 quejas?

Solución sin el ajuste por discontinuidad:

$$z = \frac{X - \mu}{\sigma} = \frac{10 - 15}{3} = -1.66$$

$$z = \frac{X - \mu}{\sigma} = \frac{14 - 15}{3} = -0.33$$

En la figura 6.29 se ilustran las circunstancias.

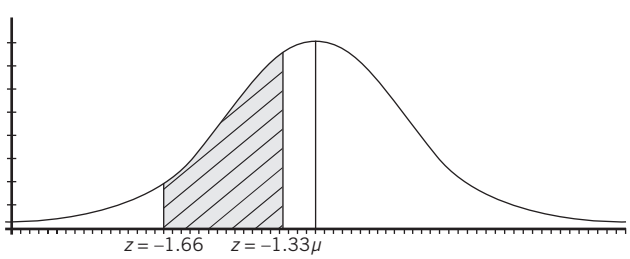


Figura 6.29 Condiciones para el ejemplo 6.16.

$$P(-0.33 \le z \le 0) = 0.1293$$
$$P(-1.66 \le z \le 0) = 0.4515$$

De donde,

$$P(-1.66 \le z \le -0.33) = 0.4515 - 0.1293 = 0.3222 = 32.22\%$$
$$P(10 \le z \le 14) = 32.22\%$$

Es decir, que hay una probabilidad de 32.22% de que en una semana se reciban entre 10 y 14 quejas.

Solución con el ajuste por discontinuidad:

$$z = \frac{X - \mu}{\sigma} = \frac{9.5 - 15}{3} = -1.83$$

$$z = \frac{X - \mu}{\sigma} = \frac{14.5 - 15}{3} = -0.17$$

Nótese que, en este ejemplo, se tomó la media de 15, sin modificación, porque es el punto de referencia para calcular ambos valores de z, ya que el área de interés, considerando que se trata de una variable discontinua, está entre 9.5, el extremo izquierdo, y 14.5, el extremo derecho.

$$P(-0.17 \le z \le 0) = 0.0675$$
$$P(-1.83 \le z \le 0) = 0.4664$$

De donde

$$P(-1.83 \le z \le -0.17) = 0.4664 - 0.0675 = 0.3989 = 39.89\%$$
$$P(10 \le X \le 14) = 39.89\%$$

Es decir, que hay una probabilidad de 39.89% de que en una semana se reciban entre 10 y 14 quejas.

Como puede verse en este ejemplo, la corrección por continuidad significó un resultado considerablemente distinto al que se obtuvo sin la corrección. Y este resultado resalta la importancia de aplicar siempre esta corrección por discontinuidad.

EJERCICIOS 6.3 Ajuste por discontinuidad

1. Se realizó una encuesta entre los alumnos de una escuela primaria para saber cuántos hermanos tenía cada uno, en promedio cada niño tiene 2 hermanos con una desviación estándar de 1. Si se elige al azar a un alumno, ¿cuál es la probabilidad de que tenga entre 3 y 4 hermanos?
2. En un hospital se reciben en promedio por día 41 consultas en el área de emergencias con una desviación estándar de 4, ¿cuál es la probabilidad de que en un día se reciban:
   a) entre 35 y 40 consultas?
   b) más de 45?
   c) menos de 30?
3. En una prueba de dominio del idioma inglés hay una media de 495 puntos con una desviación estándar de 100. Una persona desea saber la probabilidad de obtener al menos 650 puntos si presenta la prueba.
4. En un complejo de salas de cine se venden en promedio 217 canastas de palomitas grandes por día con una desviación estándar de 22; ¿cuál es la probabilidad de que en un día se vendan entre 200 y 250 canastas?
5. A la central de policía entran en promedio 27 llamadas por turno con una desviación estándar de 5, ¿cuál es la probabilidad de que en un turno se reciban 20 llamadas o menos?

6.4 Aproximación de distribuciones de probabilidad de variables discontinuas con la distribución normal

En el capítulo 5, que trata de las distribuciones discretas de probabilidad, se vio que se pueden aproximar probabilidades binomiales utilizando la distribución de Poisson. Ahora que ya se conoce la distribución normal, en las secciones siguientes se revisa cómo se puede utilizar esta distribución normal para aproximar distribuciones discretas como la binomial y la de Poisson.

6.4.1 Aproximación de la distribución binomial con la distribución normal

Fue el mismo matemático inglés Abraham de Moivre (1667-1754) quien originalmente derivó la función de densidad de la distribución normal quien, en 1733, dedujo que esta distribución normal permite aproximar los resultados de una distribución binomial con $p = 0.5$ y que la aproximación es cada vez mejor conforme n aumenta.

Posteriormente, el matemático francés Pierre Laplace (1749-1827) demostró que esa aproximación se da para cualquier probabilidad de éxito en la binomial, exceptuando 0 y 1 y, finalmente, en los años treinta, otros matemáticos demostraron que esa aproximación es verdadera también para prácticamente todas las distribuciones de probabilidad, siempre y cuando exista la desviación estándar y su valor sea diferente de 0. En el siguiente ejemplo se ilustra cómo una distribución binomial se aproxima a la distribución normal conforme p tiende a 0.5 y conforme n tiende a infinito.

EJEMPLO 6.17

Si se tiene un experimento binomial con $p = 0.3$ y $n = 3$, su distribución de probabilidad es la que se muestra en la tabla 6.3, en donde las probabilidades se calcularon con la función de probabilidad binomial que se revisó en el capítulo 5.

Tabla 6.3 Distribución de probabilidad binomial para $p = 0.3$ y $n = 3$

Núm. de éxitos de los ensayos	Probabilidad
0	0.343
1	0.441
2	0.189
3	0.027

Esta distribución de probabilidad binomial se grafica en la figura 6.30. Nótese que aunque tiene una ligera similitud con la distribución normal, es considerablemente distinta; ahora, en la tabla 6.4 se muestra la distribución de probabilidad binomial para $p = 0.4$ y $n = 20$ y en la figura 6.31 aparece la correspondiente gráfica.

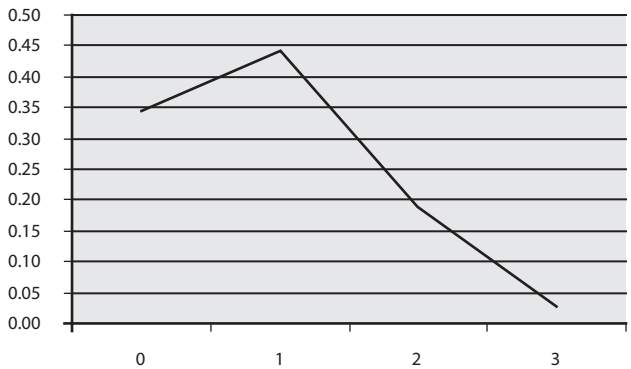


Figura 6.30 Gráfica de la distribución de probabilidad binomial para $p = 0.3$ y $n = 3$.

Tabla 6.4 La distribución de probabilidad binomial para $p = 0.4$ y $n = 20$

Núm. de éxitos de los ensayos	Probabilidad
0	0.00004
1	0.00049
2	0.00309
3	0.01235
4	0.03499
5	0.07465
6	0.12441
7	0.16588
8	0.17971
9	0.15974
10	0.11714
11	0.07099
12	0.03550
13	0.01456
14	0.00485
15	0.00129
16	0.00027
17	0.00004
18	0.00000
19	0.00000
20	0.00000

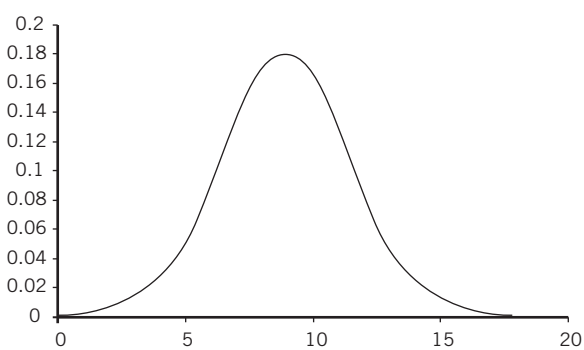


Figura 6.31 La distribución de probabilidad binomial para $p = 0.4$ y $n = 20$.

Nótese que esta gráfica con $p = 0.4$ y $n = 20$ ya es muy similar a la distribución normal, salvo que se sesga a la derecha (su cola derecha es más extendida), lo cual se debe a que la p es menor a 0.5.

Finalmente, en la tabla 6.5 se muestra la distribución de probabilidad binomial para $p = 0.5$ y $n = 20$ y en la figura 6.32 se muestra la correspondiente gráfica.

Tabla 6.5 La distribución de probabilidad binomial para $p = 0.5$ y $n = 20$

Núm. de éxitos de los ensayos	Núm. de ensayos independientes	Probabilidad de éxito de cada ensayo	Probabilidad
0	20	0.5	0.00000
1	20	0.5	0.00002
2	20	0.5	0.00018
3	20	0.5	0.00109
4	20	0.5	0.00462
5	20	0.5	0.01479
6	20	0.5	0.03696
7	20	0.5	0.07393
8	20	0.5	0.12013
9	20	0.5	0.16018
10	20	0.5	0.17620
11	20	0.5	0.16018
12	20	0.5	0.12013
13	20	0.5	0.07393
14	20	0.5	0.03696
15	20	0.5	0.01479
16	20	0.5	0.00462
17	20	0.5	0.00109
18	20	0.5	0.00018
19	20	0.5	0.00002
20	20	0.5	0.00000

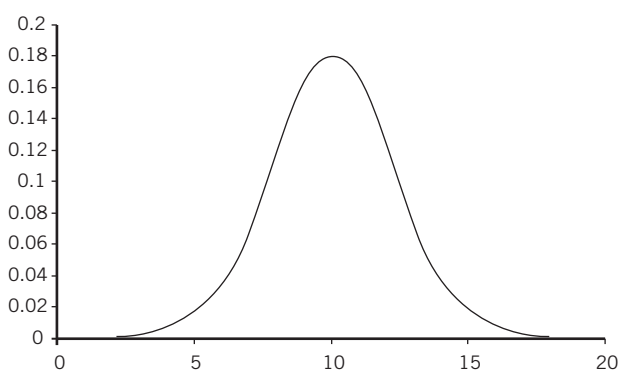


Figura 6.32 Gráfica de la distribución de probabilidad binomial para $p = 0.5$ y $n = 20$.

Estas 3 distribuciones binomiales ilustran claramente la manera en la que la distribución binomial se aproxima a la normal conforme p tiende a 0.5 y según aumenta el valor de n:

a) con $p = 0.3$ y $n = 3$
b) $p = 0.4$ y $n = 20$
c) $p = 0.5$ y $n = 20$

Existen diferentes criterios para determinar cuándo es apropiado aproximar la distribución binomial con la normal y uno de ellos es cuando: $n \geq 30$ y $np \geq 5$.

Ejemplo 6.18

En una cartera extensa de clientes se observó que 20% de los entrevistados por un representante de ventas realizan una compra, si se entrevista a 30 clientes al azar, ¿cuál es la probabilidad de que 10 o más realicen una compra?

Solución:

Se trata de una variable binomial: compra o no compra, con las siguientes características: $n = 30$, $p = 0.2$ y $q = 0.8$. Y con estos valores se pueden obtener la media y la desviación estándar de la distribución de probabilidad:

$$\mu = np = (30)(0.2) = 6$$
$$\sigma = \sqrt{npq} = \sqrt{(30)(0.2)(0.8)} = \sqrt{4.8} = 2.19$$

Representando las circunstancias en una gráfica de campana se sombrea la parte que representa la probabilidad que se busca, como en la figura 6.33.

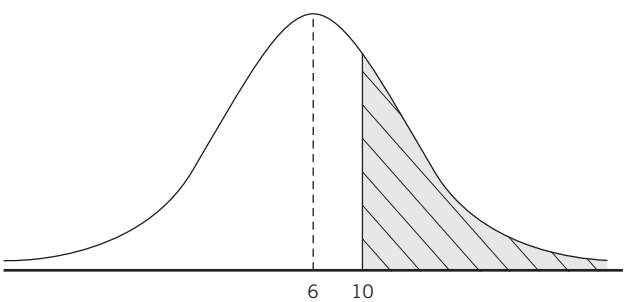


Figura 6.33 Las condiciones para el ejemplo 6.18.

Ahora, de acuerdo con los procedimientos que se utilizan para evaluar probabilidades normales y con el ajuste por discontinuidad:

$$z = \frac{X - \mu}{\sigma} = \frac{9.5 - 6}{2.19} = 1.60, \text{ y}$$

$$P(0 \le z \le 1.60) = P(6 \le X \le 10) = 0.4452$$
$$P(z \le 1.60) = P(X \ge 10) = 0.5 - 0.4452 = 0.0548$$

Con lo que se estimaría que la probabilidad de que en esa muestra de 30 clientes, 10 o más de ellos realicen una compra es de 5.48%, nótese que se aplicó la corrección de 0.5 por continuidad, ya que se utiliza la distribución normal, que es continua, para aproximar la binomial, que es discreta.

La precisión de esta estimación se puede evaluar calculando esta misma probabilidad mediante la propia distribución binomial. Para hacerlo se podría calcular la probabilidad de que compren 10, 11,..., 30 clientes o, en forma alternativa y más sencilla, restándole a 1 la suma de las probabilidades de que compren menos de 10 clientes (9, 8,..., 1), se resumen los cálculos en la tabla 6.6.

Tabla 6.6 Cálculos para el ejemplo 6.18

X	*P*(*X*)
9	0.06756361
8	0.11055863
7	0.1538207
6	0.17945748
5	0.17227918
4	0.13252245
3	0.07853182
2	0.03365649
1	0.00928455
0	0.00123794
Suma	0.93891285
1 menos la Suma	0.06108715

Así, calculando el resultado mediante la distribución binomial, la probabilidad de que compren menos de 10 clientes es:

$$P(X < 10) = 1 - 0.9389 = 0.0611$$

que es $0.0611 - 0.0548 = 0.0063$, es decir, 63 diezmilésimas menor que el resultado aproximado que se obtuvo mediante la aproximación normal, es decir, una aproximación razonable.

Ejemplo 6.19

La probabilidad de que exista un defecto en el armado de un motor de motocicleta en una planta ensambladora es de 0.15. En una muestra de 200 motores, ¿cuál es la probabilidad de que al menos 25 tengan algún tipo de defecto?

Solución:

De nuevo, se trata de una variable binomial (defecto o no defecto), con los siguientes datos: $n = 200$, $p = 0.15$ y $q = 0.85$. De donde

$$\mu = np = (200)(0.15) = 30$$
$$\sigma = \sqrt{npq} = \sqrt{(200)(0.15)(0.85)} = \sqrt{24.5} = 5.05$$

Se ilustra la situación en la figura 6.34.

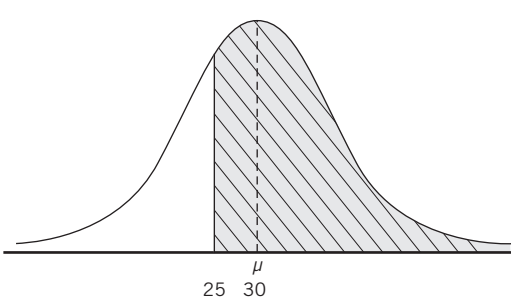


Figura 6.34 Las condiciones para el ejemplo 6.19.

$$z = \frac{X - \mu}{\sigma} = \frac{24.5 - 30}{5.05} = -1.09$$

De la tabla de áreas bajo la curva normal:

$$P(-1.09 \le z \le 0) = P(25 \le X \le 30) = 0.3621$$
$$P(z \le 1.09) = P(X \ge 25) = 0.3621 + 0.5 = 0.8621$$

Para comparar esta aproximación con el resultado exacto obtenido con la distribución binomial se tendrían que calcular las probabilidades binomiales individuales para valores de X desde 0 hasta 24 para, después, sumarlas y restar esta suma de 1. Como esto llevaría una cantidad considerable de tiempo y espacio, se adelanta aquí la solución con la función "DISTR.BINOM" que ya se usó en el capítulo anterior, la cual ahorra gran parte del trabajo. Utilizando esta función de Excel de la siguiente manera: =DISTR.BINOM(24,200,0.15,1) se obtiene como resultado 0.136817, se sabe que la probabilidad de que haya cuando mucho 24 motores con defectos es de 13.68% y, por lo tanto, la probabilidad de que cuando menos 25 de esos 200 motores tengan defectos es de 1 – 0.1368 = 0.8632, o sea, 86.32%. Este porcentaje (el exacto) es prácticamente igual al 86.21% que se obtuvo mediante la aproximación normal.

Si se comparan los resultados de este ejemplo con los del anterior se aprecia que es mejor la aproximación en este caso y esto se debe a que el tamaño de la muestra es considerablemente más grande.

6.4.2 Aproximación de la distribución de Poisson con la distribución normal

En términos generales se considera que se puede hacer una aproximación de probabilidades de Poisson cuando $\lambda \ge 10$. Para hacerlo se utilizan los parámetros de esta distribución, que son, según se vio en las secciones 5.5 y 5.6, los siguientes:

$$\mu = E(X) = \lambda = np$$

y es igual, también, a su varianza:

$$Var(X) = \lambda = np$$

por lo que la desviación estándar es:

$$\sigma = \sqrt{\lambda} = \sqrt{np}$$

Como se trata de aproximar una distribución discreta (Poisson) con la normal, que es continua, se debe aplicar el ajuste por discontinuidad, de la misma manera que se hizo con la aproximación de la binomial con la normal en la sección anterior.

EJEMPLO 6.20

En una unidad habitacional grande se recibe diariamente un promedio de 12 llamadas para mantenimiento y estas llamadas siguen una distribución de Poisson. Determine la probabilidad exacta de que en un día cualquiera se reciban menos de 5 llamadas.

Solución:

La probabilidad de que se reciban menos de 5 llamadas es la suma de las probabilidades de que se reciban 0, 1, 2, 3 o 4 llamadas. Con la distribución de Poisson:

$$P(0) = \frac{e^{-\lambda}\lambda^x}{x!} = \frac{2.71828^{-12}(12^0)}{0!}$$
$$= 2.71828^{-12} = 0.000006144$$

$$P(1) = \frac{e^{-\lambda}\lambda^x}{x!} = \frac{2.71828^{-12}(12^1)}{1!} = \frac{2.71828^{-12}(12)}{1}$$
$$= 0.000006144(12) = 0.000073$$

$$P(2) = \frac{e^{-\lambda}\lambda^x}{x!} = \frac{2.71828^{-12}(12^2)}{2!} = \frac{2.71828^{-12}(12^2)}{2}$$
$$= \frac{0.000006144(144)}{2} = 0.000442$$

$$P(3) = \frac{2.71828^{-12}(12^3)}{3!} = \frac{2.71828^{-12}(12^3)}{6}$$
$$= 0.001769472$$

$$P(4) = \frac{2.71828^{-12}(12^4)}{4!} = \frac{2.71828^{-12}(12^4)}{24}$$
$$= 0.005308418$$

De donde,

$$P(0 \le X \le 5 = 0.000006144 + 0.000073 + 0.000442 + 0.001769472 + 0.005308418 = 0.007599$$

Para aproximar ahora esta probabilidad mediante la distribución normal, en la figura 6.35 se representan las circunstancias.

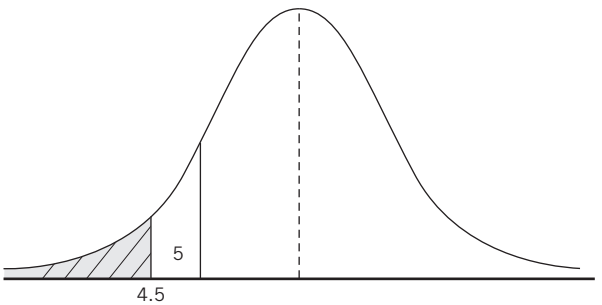


Figura 6.35 Las condiciones en el ejemplo 6.20.

La media y la desviación estándar de esta distribución de Poisson:

$$\mu = \lambda = 12$$

$$\sigma = \sqrt{\lambda} = \sqrt{12} = 3.464$$

De donde,

$$z = \frac{(5 - 0.5) - 12}{3.464} = \frac{4.5 - 12}{3.464} = \frac{-7.5}{3.464} = -2.165$$

Ahora, de la tabla de áreas bajo la curva normal,

$$P(z \le -2.165) = P(X \le 5) = 0.5 - 0.4848 = 0.0152$$

La diferencia entre los 2 resultados, aunque puede parecer grande, es de sólo 0.0152 – 0.007599 = 0.0076; es decir, 76 diezmilésimas; entonces, se puede ver que esta aproximación de probabilidades de Poisson con la normal es razonable.

Ejemplo 6.21

Si un banco recibe en promedio 5 cheques sin fondos por día, ¿cuáles son las probabilidades de que reciba 3 cheques sin fondos en un día cualquiera?

Solución:

En este caso, $\lambda = 5$ y $\sigma = \sqrt{\lambda} = \sqrt{5} = 2.236$.

a) Mediante la distribución de Poisson:

$$P(3) = \frac{e^{-\lambda}\lambda^{x}}{x!} = \frac{2.71828^{-5}(5^{3})}{3!} = \frac{0.842246}{6} = 0.14$$

La probabilidad de que el banco reciba 3 cheques sin fondos en un día cualquiera es de 0.14, o 14%.

Ahora, como se trata de aproximar esta variable discontinua con la normal, la pregunta de cuál es la probabilidad de que en un día cualquiera se reciban 3 cheques sin fondos equivale a preguntar en una distribución continua sobre la probabilidad de que el valor esté entre 2.5 y 3.5, mediante la distribución normal. En la figura 6.36 se ilustran las condiciones.

$$z = \frac{2.5 - 5}{2.236} = \frac{-2.5}{2.236} = -1.12$$

De la tabla de áreas bajo la curva normal,

$$P(-1.12 \ge z \ge 0) = 0.3686$$

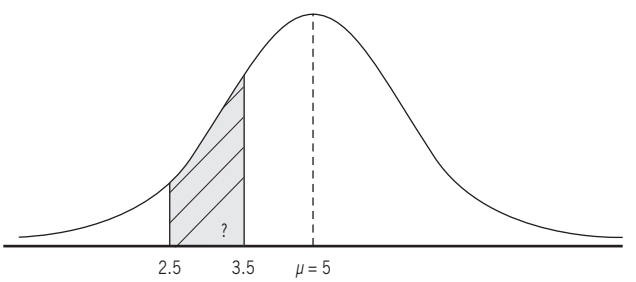


Figura 6.36 Condiciones para el ejemplo 6.21.

Por otro lado,

$$z = \frac{3.5 - 5}{2.236} = \frac{-1.5}{2.236} = -0.67$$

$$P(-0.67 \ge z \le 0) = 0.2486$$

De donde, $P(-0.12 \le z \le -0.67) = 0.3686 - 0.2486 = 0.12$.

O sea, 12% que, comparado con aquel 14% que se obtuvo mediante la distribución de Poisson, podría considerarse una aproximación razonable.

ejercicios 6.4 Aproximación de distribuciones de probabilidad de variables discontinuas con la distribución normal

Aproximación de la distribución binomial con la distribución normal

1. Se estima que 70% de las personas que entran en una tienda de ropa realizan al menos una compra. Si se toma aleatoriamente a 55 personas, ¿cuál es la probabilidad de que al menos 45 realicen una o más compras?

2. En una zapatería existe la probabilidad de 0.07 de que un artículo en particular se encuentre agotado. Si en una semana se realizan 120 pedidos:

a) ¿Cuál es la probabilidad de que 15 a 17 no puedan surtirse por falta de producto?
b) ¿Cuál es la probabilidad de que 5 a 10 no puedan surtirse?

3. En una empresa el 38% de los empleados son casados. De una muestra de 15 trabajadores, ¿cuál es la probabilidad de que:

a) más de 10 sean casados?
b) menos de 2 sean casados?
c) entre 1 y 9 sean casados?

4. Una máquina fabrica armazones para lentes y 10% de las piezas son defectuosas. En una muestra aleatoria de 100, ¿cuál es la probabilidad de que:

a) menos de 8 sean defectuosas?
b) más de 14 sean defectuosas?
c) entre 3 y 15 sean defectuosas?

5. De las llamadas que se realizan en todos los hogares del país, 28% son de larga distancia, si en un periodo se realizan 300 llamadas, ¿cuál es la probabilidad de que:

a) máximo 100 sean de larga distancia?
b) 75 o más sean de larga distancia?

Aproximación de la distribución de Poisson con la distribución normal

6. En promedio hay 13 bacterias por centímetro cuadrado en un cultivo en específico. ¿Cuál es la probabilidad de que en un segmento haya 5 o menos bacterias? Calcule utilizando:

a) La distribución de Poisson.
b) La aproximación de Poisson a la normal.

7. Cada hora en un aeropuerto aterrizan en promedio 9 aviones. ¿Cuál es la probabilidad de que en una hora aterricen entre 16 y 20 aviones? Calcule utilizando:

a) La distribución de Poisson.
b) La aproximación de Poisson a la normal.

8. Al conmutador de un centro de atención de emergencias entran en promedio 10 llamadas por minuto, calcule lo siguiente utilizando la distribución de Poisson y la aproximación de Poisson a la normal. ¿Cuál es la probabilidad de que:

a) en un minuto se reciban entre 8 y 11 llamadas?
b) entren 15 llamadas?

9. A la enfermería de una guardería ingresan en promedio 8 niños al día, calcule la probabilidad de que en un día ingresen entre 10 y 13 niños:

a) mediante distribución de Poisson.
b) por medio de la aproximación de Poisson a la normal.

10. En el detector de metales de una tienda, se detiene en promedio a 11 personas por día para una revisión. Calcule la probabilidad de que se detenga de 9 a 13 personas en un día:

a) mediante distribución de Poisson.
b) por medio de la aproximación de Poisson a la normal.

6.5 Distribución exponencial de probabilidad

Una variable aleatoria exponencial se da, por ejemplo, en el caso de un aparato electrónico que tiene la misma probabilidad de descomponerse en cualquier periodo de su vida útil (cualquier hora, cualquier día, etc.), desde nuevo y hasta el momento en el que realmente se produce la descompostura. Cuando se tiene esta probabilidad constante, se dice que hay una **tasa constante** de descompostura y el comportamiento de una variable aleatoria de este tipo sigue la distribución exponencial de probabilidad.

Tasa constante. Cuando existe la misma probabilidad de que suceda un evento en cualquier momento dado.

Otros ejemplos de variables aleatorias que siguen a esta distribución exponencial de probabilidad son:

- El tiempo que transcurre entre la llegada de llamadas a un conmutador telefónico.
- El tiempo que transcurre entre accidentes vehiculares en una intersección de calles.

- El tiempo que transcurre entre la llegada de llamadas a un teléfono de emergencias.
- La longitud que hay entre defectos en un rollo de tela.

Si una variable aleatoria se distribuye en forma exponencial con parámetro λ, su función de densidad de probabilidad es:

$$f(x) = \lambda e^{-\lambda x}, \text{ para } x > 0, \text{ y } \lambda > 0 \qquad \textbf{(6.4)}$$

La función de probabilidad acumulada se da por:

$$P(X \leq x) = 1 - e^{-\lambda x}, \text{ para } x > 0$$

La media de la distribución, μ, es: $\frac{1}{\lambda}$.
La desviación estándar de la distribución, σ, es igual a su media: $\frac{1}{\lambda}$.

En la figura 6.37 se muestran 3 distribuciones exponenciales, con $\lambda = 1$, $\lambda = 2$ y $\lambda = 3$.

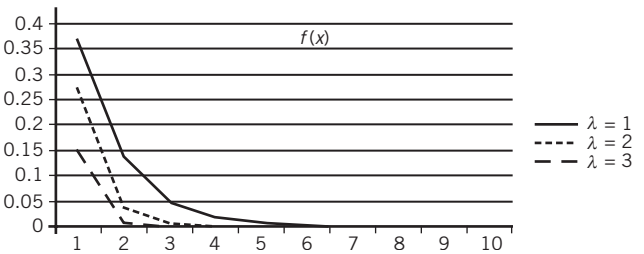


Figura 6.37. Distribuciones exponenciales, con $\lambda = 1$, $\lambda = 2$ y $\lambda = 3$, respectivamente de arriba abajo.

Ejemplo 6.22

Si una muestra de aparatos de radio enseña que los meses que transcurrieron hasta que se descompusieron fueron: 100, 115, 140, 65, 115, 150, 51, 43, 120 y 60, y, si se usa el promedio del tiempo transcurrido para estimar el parámetro λ de la producción total de estos aparatos, determine la proporción de los radios que se descompondrán en un máximo de 70 meses.

Solución:

El promedio de los meses que transcurrieron antes de las descomposturas es:

$$\bar{X} = \frac{959}{10} = 95.9$$

De donde: $\lambda = \dfrac{1}{95.9} = 0.010428$.

Y, entonces, $P(X \leq 70) = 1 - 2.71828^{-0.010428(70)} = 1 - 0.48193 = 0.5181$.

Lo cual significa que poco más de la mitad de los radios se descompondrán antes de 70 meses o, en otras palabras, que cada radio tiene una probabilidad de 51.81% de descomponerse antes de 70 meses.

Y la probabilidad de que un radio se descomponga antes de un año es:

$$P(X \leq 12) = 1 - 2.71828^{-0.010428(12)} = 1 - 0.88238 = 0.1176$$

De la misma manera, la probabilidad de que un radio se descomponga en el primer mes es:

$$P(X \leq 1) = 1 - 2.71828^{-0.010428} = 1 - 0.98962619 = 0.0103781$$

La información de este tipo puede servir para que los fabricantes fijen sus políticas de garantías para los consumidores.

Ejemplo 6.23

Un fabricante de televisores trata de determinar la duración de la garantía que debe ofrecer a sus clientes para un modelo nuevo. Las pruebas realizadas muestran que los años de vida útil de esos aparatos sigue una distribución exponencial con $\lambda = 0.10$.

a) Determine la media y la desviación estándar de la vida útil de esos televisores.
b) Si se da una garantía de 8 años, ¿qué proporción de los aparatos debe estar el fabricante preparado para reemplazar, asumiendo que el modelo propuesto es correcto?
c) Encuentre la probabilidad de que la vida útil de uno de esos televisores esté dentro del intervalo $\mu \pm 1\sigma$.

Solución:

a) $\mu = \sigma = \dfrac{1}{\lambda} = \dfrac{1}{0.10} = 10$.

b) La proporción de televisores que será necesario reemplazar se representa por los que duren de 0 a 3 años, o sea la porción oscurecida en la figura 6.38.

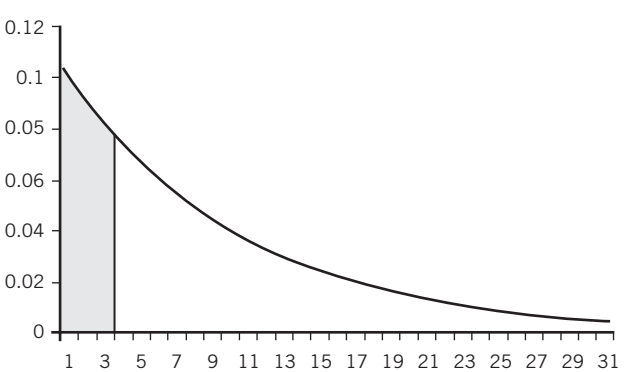


Figura 6.38 La distribución exponencial para el ejemplo 6.23.

Y, $P(X \leq 3) = 1 - 2.71828^{-0.10(3)} = 1 - 0.740819 = 0.259181$

Es decir, será necesario reemplazar 25.92% de los aparatos.

c) El intervalo $\mu \pm 1\sigma$ equivale a 10 ± 10; es decir, de 0 a 20.

Y el valor de esta probabilidad es

$$P(X \leq 20) = 1 - 2.71828^{-0.10(20)} = 1 - 0.135335 = 0.864665$$

6.5.1 Relación entre la distribución exponencial y la distribución de Poisson

Si las ocurrencias de un evento tienen una distribución de Poisson, entonces los tiempos que transcurren entre ocurrencias sucesivas tienen distribución exponencial; esto se refleja en que:

a) la media de la distribución exponencial es el inverso de la media de la distribución de Poisson,
b) la varianza de la distribución exponencial es el inverso de la desviación estándar de la distribución de Poisson.

Distribución / medida	Media aritmética	Varianza	Desviación estándar
Poisson	$\mu = E(X) = \lambda = np$	$Var(X) = \lambda = np$	$\sigma = \sqrt{\lambda}$
Exponencial	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$\frac{1}{\lambda}$

ejercicios 6.5 Distribución exponencial de probabilidad

1. Unas láminas de plástico de 50 metros tienen 5 rasgaduras en promedio producto de la máquina que las elabora. ¿Cuál es la probabilidad de que se encuentre una dentro de los primeros 5 metros?
2. Una línea de camiones que cubren la ruta México-Toluca llega a la Central de Autobuses cada 40 minutos en promedio. ¿Cuál es la probabilidad de que la diferencia entre 2 camiones sea de 70 minutos?
3. Una secretaria tiene 15 errores en un dictado compuesto de 300 palabras. ¿Cuál es la probabilidad de que:
   a) el primer error esté dentro de las primeras 10 palabras?
   b) esté en las primeras 25 palabras?
4. El departamento de urgencias recibe en promedio 5 llamadas de auxilio en una hora. ¿Cuál es la probabilidad de que la primera llamada ingrese en el transcurso de 30 minutos?
5. A un cajero automático lo utilizan 12 clientes en promedio cada hora en los periodos de más actividad en un centro comercial. ¿Cuál es la probabilidad de que pasen 10 minutos entre la llegada de 2 clientes?

6.6 Otras distribuciones de probabilidad continuas

Aparte de las distribuciones normal y exponencial, que se acaban de revisar, existen otras distribuciones continuas de probabilidad que son importantes en estadística y, en particular, para los temas que se tratan en este texto.

La distribución *t* de Student se utiliza, sobre todo, para hacer inferencias estadísticas cuando se tienen muestras pequeñas. Se revisan sus propiedades y la forma en la que se le utiliza en el capítulo 8, que se ocupa de una de las técnicas de la inferencia estadística: la estimación de parámetros.

Por su parte, la distribución χ^2 (ji cuadrada) se utiliza sobre todo en el otro tipo de inferencias estadísticas, las pruebas de hipótesis, y se aplica básicamente en 3 tipos de pruebas de hipótesis:

a) De bondad de ajuste.
b) De independencia.
c) De homogeneidad.

Se revisan sus propiedades y la forma en la que se le utiliza en el capítulo 11, que se ocupa, precisamente, de los procedimientos y aplicaciones de estos 3 tipos de pruebas de hipótesis.

Se estudia también la distribución F de Fisher, que también es continua, en diversos tipos de pruebas de hipótesis, como las que se realizan para probar la diferencia entre 2 varianzas y para las pruebas que se agrupan bajo el rubro de "análisis de varianza" y que se abordan en el capítulo 12. Se analiza con mayor detalle esta distribución en dicho capítulo 12.

6.7 Advertencia

Aunque se hizo hincapié en que la distribución normal es muy importante en estadística y si bien en este capítulo sólo se analizaron esta distribución normal y la exponencial, es importante tener presente que estas 2 distribuciones no son las únicas importantes en esta materia.

Tal como se anota en el apartado anterior, en capítulos posteriores se verán las aplicaciones de otras distribuciones continuas importantes: la distribución t de Student, la F de Fisher y la χ^2 (ji cuadrada), que se utilizan sobre todo en los temas de estimación de parámetros y pruebas de hipótesis.

Aun cuando es cierto que la distribución normal desempeña un papel preponderante en muchas aplicaciones estadísticas, también lo es que su uso indiscriminado puede conducir a conclusiones equivocadas. En el capítulo que se ocupa de pruebas de bondad del ajuste con la distribución ji cuadrada (capítulo 11) se verá una metodología que permite evaluar si una distribución observada de valores se ajusta o no a una distribución normal teórica: saber esto ayuda a decidir si el uso de la normal es adecuado o no en determinadas circunstancias.

6.8 Resumen

En este capítulo se estudiaron 2 distribuciones de probabilidad para variables continuas: la normal y la exponencial, de las cuales la normal es, con mucho, la más importante.

Se explicó cómo se usa el área bajo la curva normal para identificar probabilidades y, también, cómo se usa la distribución normal estándar, que tiene media 0 y desviación estándar de 1, para construir tablas de áreas bajo la curva normal y cómo se utilizan estas tablas para determinar:

a) las probabilidades de obtener muestras de elementos de poblaciones normales con determinadas media y desviación estándar, y la operación inversa.
b) los valores de z o de la variable de una población, conociendo el área o la probabilidad.

Se revisó también el tema de cómo calcular estas probabilidades normales haciendo un ajuste cuando se trabaja con una variable discreta o discontinua.

Por otra parte, se vieron 2 casos en los que se pueden ajustar distribuciones discretas con esta distribución normal: aproximación de probabilidades de variables binomiales y de Poisson con la distribución normal.

Finalmente se estudiaron aplicaciones de la distribución exponencial de probabilidad que tiene amplias aplicaciones en el tema de la teoría de colas o de líneas de espera.

6.9 Fórmulas del capítulo

6.2.2 Distribución normal estándar

La función de densidad de probabilidad de la distribución normal

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \qquad \textbf{(6.1)}$$

La función de densidad de probabilidad de la distribución normal estándar

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} \qquad \textbf{(6.2)}$$

6.2.4 Determinación de probabilidades para cualquier distribución normal

Estandarización de la media y la desviación estándar de cualquier distribución normal a la z de la distribución normal estándar ($\mu = 0$ y $\sigma = 1$)

$$z = \frac{X - \mu}{\sigma} \qquad \textbf{(6.3)}$$

6.2.5 Distribución exponencial de probabilidad

Función de densidad de probabilidad de la distribución exponencial

$$f(x) = \lambda e^{-\lambda x}, \text{ para } x > 0, \text{ y } \lambda > 0 \qquad \textbf{(6.4)}$$

6.10 Ejercicios adicionales

6.2.3 Tabla de áreas bajo la curva normal

Determinación del área para determinados valores de z

1. Determine el área que se encuentra entre $z = -2.75$ y $z = 1$.
2. Obtenga el área que se encuentra entre $z = -1.3$ y $z = -1.82$.
3. Precise el área que se encuentra en $z \leq -2$.
4. Especifique el área que se encuentra en $z \geq 1.97$.
5. Determine el área que se encuentra en $z \leq 0.74$.
6. Obtenga el área que se encuentra en $z \leq -1.33$ y $z \geq 1.33$.
7. Precise el área que se encuentra entre $z = -1.07$ y $z = -2.55$.
8. Especifique el área que se encuentra entre $z = -0.05$ y $z = 2.17$.
9. Determine el área que se encuentra en $z \geq -1.27$.
10. Obtenga el área que se encuentra entre $z = -1.43$ y $z = -0.33$.

6.2.3.2 Determinación de z a partir del área o la probabilidad

11. ¿Qué valor de z delimita 10% del área bajo la curva y a la derecha de la media?
12. ¿Qué valor de z delimita 5% del área bajo la curva y a la izquierda de la media?
13. ¿Cuál es el valor de z que delimita 15% del área de la curva normal en sus extremos?
14. ¿Cuál es el valor de z que delimita 30% del área de la curva normal en sus extremos?
15. ¿Qué valores de z delimitan 15% del área bajo la curva en medio de la sección derecha de la curva normal?
16. ¿Qué valores de z delimitan 20% del área bajo la curva en medio de la sección derecha de la curva normal?
17. ¿Qué valores de z delimitan 8% del área bajo la curva en medio de la sección izquierda de la curva normal?
18. ¿Qué valores de z delimitan 20% del área bajo la curva en medio de la sección izquierda de la curva normal?
19. ¿Qué valor de z delimita 5% del área la derecha de la media de una curva normal?
20. ¿Qué valor de z delimita 10.5% del área a la izquierda de la curva normal?

6.2.4 Determinación de probabilidades para cualquier distribución normal

Determinación de probabilidades o áreas

21. Un ejecutivo se dio cuenta que en los 15 países donde su empresa posee sucursales existen gastos por encima del presupuesto, en promedio cada sucursal se excede en 5 500 dólares con una desviación estándar de 800 dólares. Si estos gastos excesivos se distribuyen de manera normal, ¿qué probabilidad hay de que una sucursal se exceda por más de 4 200 dólares?
22. Un perro mediano consume en promedio 33 kilos de alimento para perro por año con una desviación estándar de 8 kilos, si el consumo sigue una distribución normal. Determine la probabilidad de que un perro consuma al año:

    a) Entre 25 y 40 kilos.
    b) Entre 33 y 46 kilos.

23. En cierta población, el consumo promedio de leche de un niño de 7 años es de 12 litros al mes con una desviación estándar de 1.4 y se distribuye de manera normal. Determine la probabilidad de que un niño cualquiera de esa edad tome 8 o menos litros al mes.
24. Un cultivo de células de piel produce en promedio 0.4 centímetros de tejido cada día, con una desviación estándar de 0.15. Determine la probabilidad de que en un día cualquiera el cultivo produzca entre 0.6 y 0.8 centímetros de tejido.
25. Una máquina tiene un tiempo de funcionamiento adecuado de 235 horas promedio sin presentar algún tipo de falla, con una desviación estándar de 22 horas. ¿Cuál es la probabilidad de que funcione de manera correcta más de 250 horas?
26. Un tipo de saborizante especial rinde en promedio 50 kilos de mezcla para chicle, con una desviación estándar de 3 kilos. Calcule la probabilidad de que rinda entre 50 y 55 kilos.
27. El consumo de watts de un hogar es de 560 diarios, en promedio, con una desviación estándar de 75. Determine la probabilidad de que en un día cualquiera se consuman entre 590 y 650 watts.
28. La vida útil de una batería de cierto modelo de teléfono celular es de 573 días, con una desviación estándar de 18. ¿Cuál es la probabilidad de que la vida útil sea de 550 o menos?
29. De una mina de carbón se obtienen en promedio 740 toneladas de producto cada 3 meses, con una desviación estándar de 27.8. Determine la probabilidad de que en un periodo se obtengan entre 735 y 750 toneladas.
30. El gasto que una empresa realiza en combustibles en el departamento de ventas es en promedio de \$12 500 al mes, con una desviación estándar de \$1 450. Determine la probabilidad de que en un mes el gasto sea de entre \$13 000 y \$14 525.

6.2.4.2 Determinación de z a partir del área o la probabilidad

31. Se sabe que la probabilidad de que se devuelvan más de 10 pistas de autos de juguete por algún defecto durante un mes es de 36%. Si el patrón de devoluciones de ese tipo de juguete sigue una distribución normal con desviación estándar de 2:

    a) Calcule la media.
    b) Obtenga la probabilidad de que en un mes se devuelvan entre 5 y 8 pistas.

32. El tiempo que se tarda un trabajador en pintar determinada superficie de la fachada de una casa tiene distribución normal y se sabe que la probabilidad de que un trabajador tarde

menos de 8 horas en hacerlo es de 5.59%. Si la desviación estándar de esos tiempos es de 2.7 horas:

a) Encuentre el promedio del tiempo que se tarda un trabajador en pintar esa superficie de fachada.
b) ¿A partir de qué tanto tiempo se tardan los trabajadores que constituyen 5% de los más lentos?

33. El tiempo que una persona tarda en leer un libro de 250 páginas sigue una distribución normal. Se sabe que la probabilidad de que una persona cualquiera se tarde cuando mucho 43 días en leer uno es de 73%, con una desviación estándar de 4. Si la distribución de estos tiempos de lectura es normal:

a) Determine el número promedio de días que los lectores de libros de estas características tardan en leer uno.
b) ¿A partir de qué número de días se encuentra el 10% que lee más rápido?

34. Un tipo de insecto puede vivir sin alimento 250 horas en promedio, con una desviación estándar de 7. ¿Entre qué valores se encuentran el 45% de los insectos alrededor de la media?

35. La cantidad de costales de harina que se rompen en los trayectos de transporte tiene una distribución normal. Si la probabilidad de que se rompan más de 18 costales de harina en uno de esos trayectos es de 2.5% y se sabe que la desviación estándar de la cantidad de costales rotos es 2:

a) Calcule la media.
b) Obtenga la probabilidad de que se rompan menos de 15 costales.

36. En una fábrica de artículos escolares se considera que el proceso de elaboración de lápices es el adecuado cuando no más de 5.6% de las piezas de los lotes de 1 000 lápices presentan algún defecto. Si la media de los defectos por lote es de 27 y la desviación estándar de 7:

a) ¿Cuál es la cantidad máxima de artículos defectuosos que debe presentar cada lote para considerar que el proceso de elaboración es el adecuado?
b) ¿Qué cantidad de lápices defectuosos presentaron 12% de lotes con más piezas con algún defecto?

37. En una evaluación estándar que aplica una empresa dedicada a la capacitación se considera que un trabajador debe obtener cuando menos 91% de aciertos en la prueba de conocimientos. Si la cantidad de aciertos sigue una distribución normal con media de 108 aciertos y una desviación estándar de 12:

a) ¿Qué cantidad mínima de aciertos debió obtener un trabajador para considerarse calificado?
b) ¿Cuál fue el máximo de aciertos de 9% de los trabajadores con el menor grado de conocimientos?

38. En una granja ovejera, los kilos de lana obtenida en cada periodo de esquila siguen una desviación normal, con una desviación estándar de 10. La probabilidad de que en un periodo de esquila se obtengan más de 110 kilos es de 10 por ciento.

a) Determine el número promedio de kilos que se obtiene en cada periodo.
b) ¿Cuál es la cantidad máxima que se obtiene dentro de 5% de periodos en que menos lana se obtiene?

6.3 Ajuste por discontinuidad

39. En una heladería se venden en promedio 47 paletas de limón a lo largo del día, con una desviación estándar de 7. ¿Cuál es la probabilidad de que en un día se vendan entre 50 y 65 paletas de limón?

40. En una central del servicio postal se reciben por día alrededor de 525 cartas con una desviación estándar de 12.

a) ¿Cuál es la probabilidad de que en un día se reciban más de 500?
b) ¿Que se reciban entre 500 y 520 cartas en un día?

41. A un taller mecánico llegan en promedio 21 autos por semana con una desviación estándar de 3.

a) ¿Cuál es la probabilidad de que en una semana lleguen entre 20 y 25 autos?
b) ¿Cuál es la probabilidad de que lleguen entre 24 y 29 autos?

42. El gerente de producción de una maquiladora determinó que la productividad al día por empleado es de 37 piezas armadas, con una desviación estándar de 4 piezas. Si considera que la productividad sigue una distribución normal, determine qué porcentaje de empleados arman entre 31 y 37 piezas en un día.

43. Un analista político se percató de que los miembros de determinado partido político se contradicen en promedio 10 veces a la semana, con una desviación estándar de 3, en sus declaraciones con la prensa. ¿Qué probabilidades hay de que en una semana cualquiera se contradigan entre 1 y 19 veces?

44. Al centro telefónico de un canal de televentas ingresan en promedio 78 llamadas por hora, con una desviación estándar de 4. Si se sabe que la entrada de llamadas sigue una distribución normal. Determine la probabilidad de que en una hora ingresen:

a) Entre 85 y 90 llamadas.
b) 70 o menos llamadas.
c) 80 o más llamadas.

45. Del portal de un grupo de música se realizan en promedio 127 descargas de su último éxito cada 30 minutos, con una desviación estándar de 18. Si las descargas siguen una distribución normal, determine la probabilidad de que en un periodo determinado de 30 minutos se descargue la canción entre 100 y 120 veces.

46. Una empresa lanzó una nueva promoción en la que se pone un premio en todas las bolsas que contienen su producto. Se observa que la máquina encargada de meter los premios no lo hace en 15 bolsas de cada lote de 150 bolsas, con una desviación estándar de 4. Determine la probabilidad de que en un lote la máquina no coloque los premios en:

a) Entre 5 y 10 bolsas.
b) Entre 13 y 27 bolsas.

47. Un fabricante de esferas de cristal notó que, en promedio, se rompen 30 piezas en cada pedido de 50 cajas, con una desviación estándar de 12. Calcule la probabilidad de que se rompan entre 20 y 35 piezas.

6.4.1 Aproximación de la distribución binomial con la distribución normal

48. Una empresa dedicada al autotransporte de pasajeros realiza corridas de la ciudad de México a Cuernavaca. La probabilidad de que una salida se retrase es de 19%, si en un día salen 102 corridas. ¿Cuál es la probabilidad de que 21 a 25 o más se retrasen?

49. Del total de jugadores que conforman la liga mexicana de futbol, 63% son mexicanos; si un equipo se conforma por 22 jugadores:

a) ¿Cuántos jugadores extranjeros aproximadamente habría?
b) ¿Cuál es la probabilidad de que 11 o más sean extranjeros?

50. Del total de niños entre 6 y 10 años, 54% se enfermarán de varicela; si en una escuela primaria hay 280 niños de estas edades:

a) ¿Cuántos aproximadamente se enfermarán de varicela?
b) ¿Cuál es la probabilidad de que menos de 170 se enfermen?

51. La probabilidad de que una persona le atine al blanco en un juego de dardos es de 0.2; si una persona tira 10 dardos, ¿cuál es la probabilidad de que acierte en por lo menos 4 ocasiones?

52. Se descubrió que un medicamento para la gripe causa algún tipo de efecto secundario en 5 de cada 100 personas que lo consumen; si se elige una muestra de 50 pacientes que ingirió dicho medicamento, ¿cuál es la probabilidad de que entre 3 y 6 presenten algún tipo de efecto secundario?

6.4.2 Aproximación de la distribución de Poisson con la distribución normal

53. Un vendedor en un estadio logra vender durante un evento 45 vasos de refresco en promedio. Calcule la probabilidad de que en un día venda 30 vasos mediante:

a) la distribución de Poisson.
b) la aproximación de Poisson con la normal.

54. En una tienda en línea se compran en promedio 17 artículos por hora. Calcule la probabilidad de que en una hora en específico se vendan 25 artículos mediante:

a) la distribución de Poisson.
b) la aproximación de Poisson con la normal.

55. El catálogo digital de una biblioteca recibe 53 consultas por hora en promedio. Calcule la probabilidad de que en una hora se reciban entre 50 y 55 consultas mediante:

a) la distribución de Poisson.
b) la aproximación de Poisson con la normal.

56. En promedio se consultan los precios de 43 artículos por hora en el verificador de precios de un supermercado. Calcule la probabilidad de que se consulten entre 50 artículos mediante:

a) la distribución de Poisson.
b) la aproximación de Poisson con la normal.

57. En promedio, una costurera termina 23 prendas en un turno de 8 horas. Calcule la probabilidad de que elabore entre 35 prendas mediante:

a) la distribución de Poisson.
b) la aproximación de Poisson con la normal.

6.5 Distribución exponencial de probabilidad

58. Un rollo de bolsas para empaque de frituras que se coloca en la línea de producción se compone de 500 bolsas y, en promedio, 30 tienen algún defecto. ¿Cuál es la probabilidad de que se encuentre un defecto dentro de las primeras 100 bolsas?

59. El tiempo de espera en la cola de un banco sigue una distribución exponencial y en promedio es de un cliente cada 10 minutos. Calcule la probabilidad de que el tiempo de espera sea menor a 5 minutos.

60. Un *campus* universitario cuenta con un servidor que conecta a internet a todos los estudiantes. Si lo utiliza un estudiante, el tiempo de respuesta promedio es de 4 segundos. Para un estudiante cualquiera, ¿cuál es la probabilidad de que transcurran a lo más 2 segundos para la llegada de la respuesta?

61. El tiempo que se tarda en tomar el pedido de un cliente en un restaurante que da servicio en su coche es de 5 minutos en promedio. ¿Qué probabilidad hay de que un cliente deba esperar 8 minutos?

62. En una estación del metro, el tiempo de llegada de trenes es de 8 minutos en promedio. Determine la probabilidad de que un usuario tenga que esperar 6 minutos.

capítulo 7

Muestreo y distribuciones muestrales

Sumario

7.1 Introducción al muestreo
- 7.1.1 Parámetros, estadísticos y estimadores
- 7.1.2 Estimación de parámetros y pruebas de hipótesis
- 7.1.3 Estimaciones por punto y estimaciones por intervalo
- 7.1.4 Muestreo aleatorio y muestreo de juicio
- 7.1.5 Muestreo aleatorio y Excel
- 7.1.6 Muestras únicas y muestras múltiples
- 7.1.7 Muestras relacionadas y muestras independientes
- 7.1.8 Tipos de muestreo aleatorio
- 7.1.9 Etapas de un estudio por muestreo
- 7.1.10 Distribuciones muestrales

7.2 Distribución muestral de la media
- 7.2.1 Desarrollo
- 7.2.2 Tres conclusiones importantes que se desprenden de la distribución muestral de la media: el teorema central del límite
- 7.2.3 Fórmula del error estándar de la media y factor de corrección por población finita
- 7.2.4 Consideraciones adicionales sobre la distribución muestral de la media
- 7.2.5 Aplicaciones del análisis de la distribución muestral de la media

7.3 Distribución muestral de la proporción
- 7.3.1 Desarrollo
- 7.3.2 Tres conclusiones importantes sobre la distribución muestral de la proporción
- 7.3.3 Fórmula del error estándar de la proporción y factor de corrección por población finita
- 7.3.4 Consideraciones adicionales sobre la distribución muestral de la proporción

7.4 Distribución muestral de la varianza
- 7.4.1 Distribuciones muestrales sin reemplazo y con reemplazo
- 7.4.2 Estimadores insesgados y estimadores sesgados

7.5 Resumen

7.6 Fórmulas del capítulo

7.7 Ejercicios adicionales

Como se comentó antes, una *población* es el conjunto de todos los elementos o unidades que son de interés para un estudio determinado y una *muestra* es un subconjunto de los elementos de la población. La principal característica que debe tener una muestra estadística es que sea *representativa* de la población de la cual se extrae, porque el principal propósito de la obtención de muestras consiste en hacer inferencias sobre la población correspondiente y esta labor se realiza primordialmente a través de 2 técnicas: la estimación de parámetros y las pruebas de hipótesis.

El muestreo se utiliza cuando resulta más fácil, más rápido o más económico estudiar una parte de la población y no la población completa. En muchas ocasiones estas mayores sencillez, velocidad y economía dan como resultado conclusiones que son igualmente útiles que las que se obtendrían de estudiarse la población completa. Además, se utiliza también el muestreo en los casos en que las restricciones de tiempo, dinero o complejidad hacen que resulte imposible estudiar la población completa.

En este capítulo se revisan diversos conceptos importantes relacionados con el muestreo y se revisa también con detalle el tema de las distribuciones muestrales que, aunque es primordialmente teórico, es la base de donde se desprenden los mecanismos con que se hacen estimaciones de parámetros y pruebas de hipótesis.

7.1 Introducción al muestreo

En esta sección se repasan diversos temas básicos para comprender las útiles herramientas de la estadística inferencial y sus múltiples aplicaciones en las áreas de la administración, la economía y, en general, las ciencias sociales.

Se revisan los importantes conceptos de parámetros, estadísticos y estimadores y se explican cuáles son las 2 principales técnicas de la inferencia estadística: la estimación de parámetros y las pruebas de hipótesis.

Asimismo, se analizan brevemente otros conceptos relevantes para comprender los diferentes tipos de muestreo y se termina por introducir el tema del capítulo, las distribuciones muestrales, que son la base teórica de las técnicas de muestreo que se aplican en la práctica.

7.1.1 Parámetros, estadísticos y estimadores

Como se vio en capítulos anteriores, los estudios estadísticos, sean muestrales o no, se hacen con base en la descripción de conjuntos de datos mediante gráficas, tablas o medidas que permiten apreciar las características de los datos. Esta descripción convierte a los datos en información útil para la toma de decisiones ya que permite evaluar de mejor manera las circunstancias del caso en estudio. Cuando se realizan estudios por muestreo, las medidas que más comúnmente se manejan son medias (promedios aritméticos) y proporciones. Así, un ejemplo de una inferencia estadística podría decir algo así como: "se estima que el *promedio* de los ingresos mensuales de la población en estudio está entre $600 y $650, y se tiene una confianza de 99% de que esta afirmación es correcta"; otro ejemplo: "se puede afirmar, con una probabilidad de 95% de estar en lo correcto, que la *proporción* de consumidores de esta población que prefieren nuestro producto está entre 35 y 40%". En el primer ejemplo se estima la media de la población con base en la media de la muestra y, en el segundo, la proporción de la población con base en la proporción de la muestra.

Al observar estos ejemplos resulta evidente que, para analizar la teoría y las aplicaciones del muestreo, es necesario marcar la diferencia entre la medida de la muestra y la correspondiente medida de la población. Así, a las medidas de las poblaciones se les denomina **parámetros poblacionales** y a las medidas de las muestras se les denomina **estadísticos muestrales**. Se les denomina, también, en términos más breves, *parámetros* y *estadísticos*. Además, como se utilizan símbolos para representar estas diversas medidas, también se manejan diferentes símbolos para representarlas:

Parámetros poblacionales. Medidas de las poblaciones.
Estadísticos muestrales. Medidas de las muestras.

Medida	Parámetro poblacional	Estadístico muestral
Media aritmética	μ	$\overline{X}$
Proporción	π	P
Total	T	T
Varianza	σ^2	s^2
Desviación estándar	σ	S
Núm. de elementos	N	n

7.1.2 Estimación de parámetros y pruebas de hipótesis

Tal como se mencionó antes, la estimación de parámetros (de la que se dieron varios ejemplos en secciones anteriores) son, junto con las pruebas de hipótesis, las 2 técnicas que más comúnmente se utilizan para hacer inferencias sobre los parámetros de las poblaciones de interés.

Abundando sobre la terminología de las estimaciones de parámetros, cuando se utiliza un estadístico para estimar un parámetro se dice que el estadístico se convierte en un **estimador**.

Estimador. Cuando se utiliza un estadístico para estimar un parámetro se dice que el estadístico se convierte en un estimador.
Pruebas de hipótesis. Utilizar datos muestrales para evaluar la posible veracidad o precisión de suposiciones que se hacen sobre la población de interés.

Por su parte, las **pruebas de hipótesis** consisten en utilizar datos muestrales para evaluar la posible veracidad o precisión de suposiciones que se hacen sobre la población de interés. Por ejemplo, en una empresa manufacturera se podría tener el caso de que, para poder considerar que el proceso de fabricación de botellas funciona bien, no se debe tener una proporción de botellas defectuosas que sea superior a 1%. Una prueba de hipótesis consistiría en probar, precisamente, la hipótesis (suposición) de que el proceso está bajo control y que, por lo tanto, la proporción de botellas defectuosas que se fabrican no es mayor de 1% (el parámetro de la población); para probar esta hipótesis se obtendría una muestra de botellas y se determinaría la proporción de botellas defectuosas (el estadístico muestral); en términos generales, el procedimiento de prueba consiste en decidir:

1. Si los datos de la muestra *coinciden* con el parámetro de la población (la proporción de botellas defectuosas de la muestra es muy cercana a 1%), se acepta la hipótesis de que el proceso está bajo control.
2. Si los datos de la muestra *no coinciden* con el parámetro de la población (la proporción de botellas defectuosas es muy superior a 1%), se concluye que la hipótesis de que el proceso está bajo control no es cierta.

Si, como en el caso anterior, la conclusión de la prueba de hipótesis es que el proceso no opera de acuerdo con lo esperado, se pueden tomar medidas para corregir la situación. La mayor parte de lo que resta de este libro se relaciona con estas 2 técnicas. En las secciones siguientes y en los demás capítulos se revisan diversas técnicas y casos de aplicación del muestreo en la práctica.

7.1.3 Estimaciones por punto y estimaciones por intervalo

La estimación de parámetros se ejemplificó en la sección anterior, cuando se dieron ilustraciones sobre el uso de un estadístico para *estimar* un parámetro. Otro ejemplo: "se considera, con una probabilidad de 99% de que la estimación sea correcta, que la proporción de electores que votarían en favor del candidato X está entre 30 y 35%". Esta afirmación es un ejemplo de una *estimación por intervalo*. La razón del nombre es clara: se estima el parámetro de la población por medio de un intervalo especificado por 2 valores, a los que se denomina el *límite superior del intervalo* (35%) y el *límite inferior del intervalo* (30%).

Una **estimación por punto** es la que se realiza utilizando un solo número para la estimación. De hecho, al obtener una muestra, el estadístico muestral representa una estimación puntual del correspondiente parámetro. Así, al obtener una muestra y calcular su media, se puede utilizar esta media muestral como estimador por punto, o puntual, de la media de la correspondiente población.

Estimación por punto. Se realiza utilizando un solo número para la estimación.

La principal ventaja de las estimaciones por intervalo, y que es al mismo tiempo la razón por la cual son las más ampliamente usadas, es que se les pueden asociar juicios de probabilidad respecto a la confianza de estar en lo correcto al hacer la estimación. Como se verá después, dado el hecho de que las estimaciones se hacen las más de las veces con base en la distribución normal de probabilidad, no es posible establecer la probabilidad de ocurrencia de un único valor en una escala continua, sólo se puede establecer la probabilidad de que un valor determinado caiga dentro de un intervalo.

7.1.4 Muestreo aleatorio y muestreo de juicio

Para que las inferencias que se hacen con base en muestras sean útiles, la muestra que se utiliza debe ser *representativa* de la población de la que se extrae; por ello, las muestras que se utilizan en estadística deben ser muestras aleatorias; es decir, deben ser muestras elegidas mediante métodos al azar. En otras palabras, una **muestra aleatoria** es aquella que se elige de manera que se conoce la probabilidad de elegir a cada uno de sus elementos y el caso más común es cuando todos y cada uno de los elementos de la población tienen la misma probabilidad de ser elegidos para la muestra. Esto es así porque se requiere la aleatoriedad para establecer los criterios de probabilidad que se asocian a los resultados del muestreo, es decir, las inferencias.

Muestra aleatoria. Aquélla en que todos los elementos de la población tienen o una probabilidad conocida de aparecer en una muestra o la misma para salir en ellas.

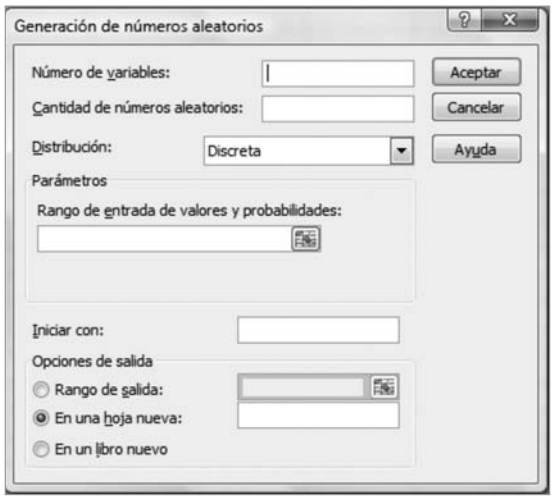
Se puede ilustrar la obtención de una muestra aleatoria con el caso en el que se dispone de un listado numerado de todos los elementos de la población y, después de determinar el número de elementos de la muestra, se procede a su conformación mediante el azar, utilizando una tómbola con números o una tabla de números aleatorios. Por ejemplo, si se tuviera una población de $N = 1\,000$ elementos y se deseara tomar una muestra de $n = 100$ elementos, se usaría una tómbola que tuviera esferas numeradas del 0 al 9 y se sacarían tríos (tercias) de esferas que identificarían números del 000 al 999, correspondientes a los números de identificación de cada uno de los elementos de la población, hasta llegar a tener los 100 elementos que se desean para la muestra.

La obtención de la muestra anterior con una tabla de números aleatorios, que son tablas que contienen números a los que propiamente se llama *pseudoaleatorios* y que son conjuntos de números generados al azar por computadora, se llevaría a cabo partiendo al azar desde cualquier punto (número) de la tabla y desplazándose también en cualquier dirección en la tabla, tomando grupos de números de 3 en 3 que, al igual que con la tómbola, corresponderían a los números de identificación de los elementos de la población.

Aunque también existen métodos de muestreo para muestras aleatorias en las que no todos los elementos de la población tienen la misma probabilidad de incluirse en la muestra (pero sí una probabilidad conocida) y para muestras que no se escogen al azar (*muestras de juicio*), en el presente texto no se les analiza.

7.1.5 Muestreo aleatorio y Excel

El complemento "Análisis de datos" de Excel cuenta con 2 herramientas relacionadas con la extracción de una muestra aleatoria: generación de números aleatorios y muestra. En los párrafos siguientes se explica cómo se usa cada uno de ellos.



7.1.5.1 Generación de números aleatorios

Este mecanismo genera un conjunto de números aleatorios, cuando se elige esta opción a partir de la sección "Análisis de datos" en la pestaña "Datos" de la cinta de opciones de Excel, aparece el cuadro de diálogo mostrado a la izquierda.

El número que se anote en el cuadro "Número de variables" indica cuántas columnas de valores se desean incluir en la tabla de resultados. Si no se establece nada, Excel rellena todas las columnas especificadas en el rango de salida.

La cantidad de números aleatorios permite determinar el número de puntos de datos que se desean ver, igual que con el anterior, si no se especifica nada aquí, Excel rellena todas las columnas establecidas en el rango de salida.

En el cuadro de distribución se puede elegir entre las siguientes distribuciones: discreta, uniforme, normal, Bernoulli, binomial, Poisson y frecuencia relativa, y determina el tipo de distribución que utiliza Excel para generar los números aleatorios. En la tabla 7.1 se lista cada una de estas distribuciones, junto con el conjunto de parámetros en los que se basan, anotados en la segunda columna.

Tabla 7.1 Distribuciones en las que se basa la generación de números aleatorios con Excel, sus parámetros y los valores utilizados para el ejemplo

Distribución	Parámetros	Valores usados en el ejemplo
Uniforme	Entre 0 y 1	
Normal	Media 0, desviación estándar 1	
Bernoulli	Solicita la probabilidad de la distribución, que puede ser cualquier valor entre 0 y 1	$P = 0.1$
Binomial	Solicita la probabilidad de la distribución, que puede ser cualquier valor entre 0 y 1 y pide, además, el número de muestras.	$P = 0.1$; núm. de muestras, 100.
Poisson	Solicita lambda, λ, la media de la distribución Poisson.	$\lambda = 5$
Frecuencia relativa	Aquí pide un rango: "De *x* a *y*", y solicita también el incremento y pide el número de repeticiones de números y de secuencias.	De 0 a 100 con incrementos de 5. Repeticiones de números y de secuencias de 1 en ambos casos.
Discreta	Solicita un rango en donde se encuentra un conjunto de valores y probabilidades; es decir, una distribución de frecuencias específica.	No se usó.

En la segunda columna se muestran los parámetros utilizados para generar las series de números aleatorios que se presentan en la tabla 7.2. En todos los casos se solicitaron 20 números aleatorios.

Tabla 7.2 Números aleatorios generados a partir de diferentes distribuciones

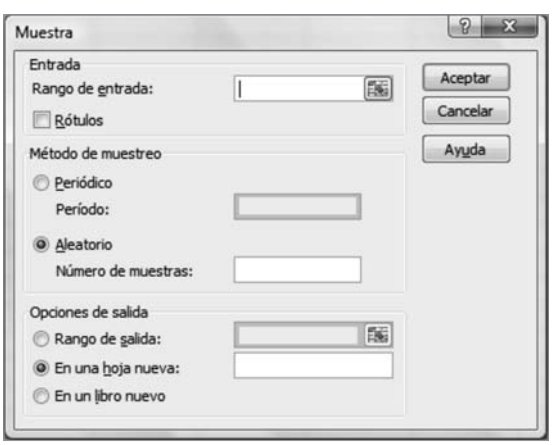
Uniforme	Normal	Bernoulli	Binomial	Poisson	Frecuencia relativa	Uniforme × 1 000
0.82	2.64	0	9	2	0	820
0.70	–1.59	0	7	7	5	700
0.39	0.75	0	7	3	10	390
0.06	1.47	0	10	11	15	60
0.22	–1.92	0	13	3	20	220
0.74	–1.52	0	4	5	25	740
0.22	–0.33	0	9	4	30	220
0.96	–0.06	0	15	3	35	960

Uniforme	Normal	Bernoulli	Binomial	Poisson	Frecuencia relativa	Uniforme × 1 000
0.71	0.68	0	9	5	40	710
0.58	0.32	1	7	2	45	580
0.01	1.45	0	14	5	50	10
0.95	0.50	1	5	4	55	950
0.94	0.33	0	8	4	60	940
0.49	1.54	0	12	3	65	490
0.86	0.42	0	9	2	70	860
0.38	2.59	1	10	6	75	380
0.36	0.35	0	15	4	80	360
0.03	0.90	0	10	5	85	30
0.92	–0.41	0	6	4	90	920
1.00	1.51	0	11	4	95	1 000
					100	

La distribución uniforme es la que puede utilizarse para generar una lista de números aleatorios. Siguiendo con el ejemplo de la población cuyos elementos se numeran del 1 al 1 000, si se quisiera extraer una muestra aleatoria de $n = 20$ elementos, se utilizarían los números obtenidos en la tabla 7.2 y, multiplicados por 1 000, se obtendrían los números dentro de la numeración de los elementos de la población, con lo que se generarían los de la última columna de esta tabla que marcan, precisamente, los números de los elementos de la población que se incluirían en la muestra.

Como puede verse, la generación de números aleatorios con Excel es muy versátil. Aquí sólo se desea destacar que los "números aleatorios", generados a partir de la frecuencia relativa, en realidad no son aleatorios sino que son una distribución de frecuencias generada a partir de los parámetros especificados.

7.1.5.2 Muestra



Al elegir esta opción aparece el cuadro de diálogo que se presenta a la derecha.

Esta herramienta crea una muestra aleatoria o una *muestra periódica* (o *sistemática*, como se le llama aquí) a partir de los elementos de la población especificados en el "rango de entrada".

Para generar una muestra aleatoria de 20 elementos a partir de la población de 1 000 elementos, se pueden llenar las celdas A1 hasta la A1 000 con los números del 1 al 1 000 y después marcar ese rango de celdas como el "rango de entrada" A1:A1 000, elegir el "método de muestreo" "aleatorio" y un "rango de salida" de 20 celdas, por ejemplo C1:C20. Haciendo esto se obtuvieron los números siguientes:

922
444
397
854
311
768
225
682
139
596
53

510
967
61
592
842
299
756
213
670

Que serían, entonces, los elementos de la población que se incluirían en la muestra.

7.1.6 Muestras únicas y muestras múltiples

Muestra aleatoria simple. Es una sola muestra aleatoria.
Muestras múltiples. Es cuando se obtiene más de una muestra para el estudio.

Cuando se planea un estudio por muestreo es necesario determinar el *diseño de la muestra*. Se puede revisar este diseño desde 2 puntos de vista: el número de muestras y el procedimiento utilizado para elegir los elementos de la muestra o muestras.

Respecto al *número de las muestras* se dice que se tiene una **muestra aleatoria simple** cuando se obtiene *una* sola muestra aleatoria. Se dice que se tienen **muestras múltiples** cuando se obtiene más de una muestra para el estudio.

7.1.7 Muestras relacionadas y muestras independientes

El caso más común de muestras múltiples es cuando se eligen 2 muestras. En estas ocasiones se pueden extraer las 2 muestras de una sola población cuando se pretende, por ejemplo, probar los efectos de un curso de capacitación. En este caso, se obtiene la muestra de personas que participan en el curso de capacitación y se mide *antes* y *después* del curso. Aquí se dice que se tienen **muestras relacionadas** o *muestras no independientes*, puesto que las mediciones en las 2 muestras son en realidad 2 mediciones a la misma muestra, aunque en condiciones diferentes; cuando las muestras se obtienen de poblaciones distintas se dice que se trata de **muestras independientes**. Un ejemplo de esto es cuando se desea comparar los niveles de gastos en ropa en 2 poblaciones (como, por ejemplo, tomando muestras en ciudades distintas), como se verá en su momento, los métodos estadísticos para manejar muestras independientes y muestras relacionadas son diferentes.

Muestras relacionadas. Son mediciones diferentes de la misma muestra pero en condiciones diferentes.
Muestras independientes. Muestras que se obtienen de poblaciones distintas.

7.1.8 Tipos de muestreo aleatorio

En el caso del muestreo al azar, los principales *métodos de muestreo* son:

1. *Muestreo aleatorio simple*. En éste se selecciona una sola muestra y todos los elementos de la población tienen la misma probabilidad de ser escogidos para la muestra.
2. *Muestreo aleatorio sistemático*. En este método de muestreo se cuenta, por lo general, con una lista que contiene todos los elementos de la población ordenados según algún criterio como orden alfabético (una lista de empleados), fecha de la transacción (facturas de venta, cheques expedidos) o algún otro, si el orden en el que se encuentran los elementos no tiene relación con las características que se desea estudiar, entonces se puede aplicar el muestreo sistemático para elegir la muestra aleatoria. El procedimiento consistiría en elegir cada k-ésimo elemento de la lista para que forme parte de la muestra. Por ejemplo, si se tiene una lista de 1 000 elementos de determinada población y se desea formar una muestra de 100 de ellos, se elegiría cada décimo elemento de la lista como parte de la muestra, sorteando el primero de ellos; en otras palabras, se podría sortear fácilmente en cuál de los primeros 10 elementos se debe comenzar y a partir de allí, elegir para la muestra cada décimo elemento; si se comenzara con el elemento 4 el segundo elemento de la muestra sería el que ocupe el lugar 14 de la lista, el tercero sería el número 24, y así sucesivamente.
3. *Muestreo aleatorio estratificado*. En este tipo de muestreo se divide la población en subconjuntos —normalmente unos pocos— a los que se denomina *estratos*, para después sacar una muestra aleatoria simple en cada uno de estos estratos. Al obtener los resultados de cada estrato se les combina para sacar una estimación combinada para la población como un todo. Se aplica este esquema de muestreo cuando la variabilidad (desviación estándar) de la variable de interés es menor dentro de cada uno de los estratos que en la población como un todo; si se cumple esto, las estimaciones que se calculan mediante muestreo aleatorio estratificado son más precisas que las que se obtienen utilizando muestras aleatorias simples en la misma población debido a que, al ser más homogéneas las mediciones dentro de cada uno de los estratos que en la población total, se reduce el error muestral de las estimaciones (la desviación estándar de los estimadores).

   Cuando se utiliza el muestreo aleatorio estratificado, es preferible hacer que los tamaños de las muestras en cada estrato sean proporcionales a los tamaños de los estratos respecto a la población; también es posible hacer inferencias con base en muestras que no sean proporcionales a los tamaños de sus correspondientes estratos, aunque esto complica un poco la labor de ponderar (asignar pesos o importancias relativas) a los resultados de cada una de las muestras.

Un ejemplo de un estudio que pudiera convenir realizar mediante muestreo estratificado sería el estudio del promedio de ingresos de una población heterogénea; se podría estratificar la población en clases baja, media y alta, de esta manera, la dispersión de la variable de ingresos sería menor en cada uno de los estratos que en la población completa.

4. *Muestreo por conglomerados*. Aquí se divide a la población en subgrupos a los que se denomina **conglomerados**, y que son relativamente pequeños (con pocos elementos) y numerosos, para después obtener una muestra aleatoria de conglomerados. El criterio que se sigue para conformar los conglomerados es que sus elementos sean similares entre sí en términos de las características de interés. Este método de muestreo resulta especialmente útil cuando no se dispone de un listado de todos los elementos de la población o cuando estos elementos están separados físicamente entre sí por distancias considerables. Como se agrupa a los elementos de los conglomerados de acuerdo con su homogeneidad, es más conveniente tener muchos conglomerados pequeños que tener pocos conglomerados grandes porque, con la primera condición, se reduce el riesgo de dejar fuera elementos con determinadas características.

   Un ejemplo de un estudio en el que convendría el muestreo por conglomerados sería cuando se desea estudiar una población conformada por las familias de un área urbana grande. En este caso es prácticamente imposible obtener un listado de todas las familias involucradas pero, si se puede obtener un plano detallado y completo de la zona, se podría considerar que las manzanas del plano son los conglomerados y elegir una muestra aleatoria de ellos para conformar la muestra.

Conglomerados. Subgrupos de una población que son relativamente pequeños (con pocos elementos) y numerosos.

7.1.9 Etapas de un estudio por muestreo

A continuación se explican brevemente los pasos que es conveniente seguir para realizar un estudio por muestreo:

1. Definir los objetivos de la investigación. Ésta es la etapa inicial de cualquier estudio científico.
2. Definir la población a estudiar. La mejor manera de definir esta población consiste en especificar cuáles son todos y cada uno de sus elementos. Si es posible hacer esto, se tiene lo que se conoce como **marco muestral** y que en su forma ideal significa contar con un listado que contenga todas las *unidades de muestreo*. Se entiende por **unidad de muestreo** a cada uno de los elementos de la población que se va a estudiar; estos elementos pudieran ser familias, personas, casas habitación, facturas, cheques, animales, piezas fabricadas, etcétera.
3. Elegir el diseño de la muestra. Como se mencionó en la sección anterior se debe decidir si se tomará una sola muestra o varias, en el caso de ser varias de ellas se debe especificar si serán muestras independientes o muestras relacionadas. Finalmente, hay que determinar si se utilizará el muestreo aleatorio simple, el estratificado, el sistemático, el que se basa en conglomerados, o algún otro tipo.
4. Realizar una prueba piloto. En ocasiones es conveniente, o hasta imprescindible, realizar un estudio en pequeña escala para determinar si el diseño de la muestra no presenta dificultades. Esto es especialmente cierto para estudios que se realizan por primera vez o en estudios en los que se aplican encuestas (en este último caso, además de probar el diseño de la muestra se prueba el diseño del cuestionario que se utiliza). En estudios que se repiten, con frecuencia no es necesario realizar este paso.
5. Recolección, procesamiento y análisis de la información.
6. Presentación de conclusiones.

Marco muestral. La especificación de todos y cada uno de los elementos de la población a estudiar.
Unidad de muestreo. Cada uno de los elementos a estudiar de una población.

Como el presente libro es un texto introductorio, se hará hincapié en las etapas de análisis de la información y de presentación de conclusiones. Sin embargo, a lo largo del material se harán observaciones que abundan sobre los detalles de todas estas etapas.

7.1.10 Distribuciones muestrales

El tema de las distribuciones muestrales es de vital importancia para la estadística inferencial, ya que, como veremos más adelante, las conclusiones a las que conduce son el fundamento de las técnicas de estimación de parámetros y de pruebas de hipótesis que son, a su vez, los pilares de las inferencias con muestras aleatorias. La manera más sencilla de comprender los conceptos relacionados con el tema es a través de ejemplos ilustrativos; sin embargo, antes de pasar a los ejemplos, conviene plantear un panorama general.

Distribución muestral. El conjunto de todas las muestras distintas de determinado tamaño *n* que es posible extraer de una población de tamaño *N*.

Una **distribución muestral** es el conjunto de todas las muestras distintas de determinado tamaño *n* que es posible extraer de una población de tamaño *N*. Conviene analizar algunos puntos en la definición anterior. En primer lugar, se debe observar que se trata del conjunto de *todas* las muestras distintas que es posible extraer de determinada población; al analizar este conjunto exhaustivo se extraen conclusiones respecto al posible comportamiento de *una sola* muestra; en otras palabras, lo que interesa es una sola muestra, no todas ellas, pero analizando ese conjunto de todas las que es posible extraer, se puede conocer el posible comportamiento de la muestra única que interesa.

Con base en la definición anterior de distribución muestral, se puede hablar de la distribución muestral de un estadístico y en las secciones siguientes se analizan las 3 más importantes: la distribución muestral de la media, la distribución muestral de la proporción y la distribución muestral de la varianza.

Para terminar esta sección, y a manera de ejemplo, diremos que la distribución muestral de la media es el conjunto de las medias de todas las muestras distintas de tamaño *n* que es posible extraer de una población de tamaño *N*. De manera paralela aunque más abreviada, la distribución muestral de la proporción es el conjunto de las proporciones de todas las muestras que es posible extraer de determinada población.

ejercicios 7.1 Introducción al muestreo

1. ¿Qué es una muestra?
2. Explique cada uno de los siguientes términos:

a) Estadístico.
b) Parámetro.
c) Estimador.

3. ¿Por qué es útil el muestreo?
4. ¿Cuál es la diferencia entre un estimador por punto y un estimador por intervalo?
5. ¿Qué es una muestra aleatoria?
6. Explique en qué consisten los siguientes métodos para obtener muestras aleatorias:

a) Muestreo simple.
b) Muestreo estratificado.
c) Muestreo sistemático.
d) Muestreo por conglomerados.

7. ¿Cuándo se dice que 2 muestras son independientes?
8. Explique las etapas de un estudio por muestreo.
9. ¿Cuál es la importancia práctica del análisis de las distribuciones muestrales de estadísticos?

7.2 Distribución muestral de la media

Aunque el análisis que sigue de las distribuciones muestrales es principalmente teórico, es de vital importancia, ya que las conclusiones que de él se derivan son el fundamento de las 2 técnicas básicas de la inferencia estadística: la estimación de parámetros y las pruebas de hipótesis. Se recomienda prestar atención al desarrollo de las exposiciones de las diversas distribuciones muestrales que se presentan para poder comprender cabalmente las conclusiones que se obtienen del análisis. Estas conclusiones, dada su importancia, se enfatizan en secciones separadas.

Si se revisa el contenido del capítulo se podrá ver que se analizan 3 distribuciones muestrales: de la media, de la proporción y de la varianza. Cada una de ellas se relaciona directamente con estimaciones de parámetros y pruebas de hipótesis sobre el parámetro correspondiente. Así, del análisis de la distribución muestral de la media se desprenden los mecanismos que permiten hacer esas inferencias estadísticas sobre, precisamente, la media, e igual sucede para las otras 2: del análisis de la distribución muestral de la proporción se desprenden los mecanismos que permiten hacer inferencias estadísticas sobre la proporción, aplicando lo mismo para la distribución muestral de la varianza.

Distribución muestral de la media. Conjunto de las medias de todas las muestras de tamaño *n* que es posible obtener de una población de tamaño *N*.

Tal como se mencionó en la sección anterior, la **distribución muestral de la media** es el conjunto de las medias de todas las muestras de tamaño *n* que es posible obtener de una población de tamaño *N*. En esta sección se revisa el comportamiento de las distribuciones muestrales de la media porque, como ya se explicó, de este comportamiento se desprenden los procedimientos para aplicar los 2 principales mecanismos de la inferencia estadística, la estimación de parámetros y las pruebas de hipótesis sobre medias, que a su vez representan una porción importante de las aplicaciones que se revisarán en buena parte de los capítulos restantes de este libro.

7.2.1 Desarrollo

Antes de pasar a un ejemplo para ilustrar esta distribución muestral, y considerando que se trata de un ejemplo un tanto extenso, conviene visualizar antes lo que se hará:

- Se supondrá una población con N elementos.
- Se calcularán la media aritmética y la desviación estándar de la población.
- Se determinará el número total de muestras distintas de tamaño n que es posible extraer y se concluirá cuáles son; es decir, se enumerará la distribución muestral (el conjunto de todas las muestras de ese tamaño que es posible obtener de esa población).
- Se resolverá la media de cada una de las muestras identificadas en el paso 3; con esto ya se tiene la distribución muestral de la media (el conjunto de las medias de todas las muestras de ese tamaño que es posible obtener de la población).
- Se calcularán la media y la desviación estándar de la distribución muestral de la media.
- Se obtendrán las conclusiones importantes para el muestreo al comparar los parámetros poblacionales con los valores obtenidos para la distribución muestral de la media.
- Se observará también la forma de la distribución muestral de la media para obtener otra importante conclusión: la que permite utilizar la distribución normal para hacer los juicios sobre la probabilidad de estar en lo correcto cuando se hacen inferencias.

El ejemplo que se presenta en seguida supone una población con sólo 5 elementos ($N = 5$) y una muestra de tamaño $n = 2$. Esto es una situación extremadamente simplificada que tiene como propósito facilitar la explicación de los conceptos importantes del tema, mas estos conceptos son igualmente aplicables a poblaciones muy numerosas y muestras aún mayores que se manejan en la práctica.

Ejemplo 7.1

Suponga que se tiene una población de 5 familias ($N = 5$) y la variable que se estudia es el número de hijos de cada familia. Los datos correspondientes aparecen en las 2 primeras columnas de la tabla 7.3.

Tabla 7.3 Número de hijos

Familia	Hijos X	$(X - \mu)$	$(X - \mu)^2$
Pérez	2	– 4	16
Gómez	4	– 2	4
Durán	6	0	0
Hidalgo	8	2	4
Juárez	10	4	16
Totales	30	0	40

Se determinan ahora la media aritmética y la desviación estándar de esta población. Los cálculos necesarios aparecen en la tabla 7.3. La media:

$$\mu = \frac{\sum X}{N} = \frac{30}{5} = 6$$

La desviación estándar:

$$\sigma = \sqrt{\frac{\sum (X - \mu)^2}{N}} = \sqrt{\frac{40}{5}} = \sqrt{8} = 2.8284271$$

Con un tamaño de muestra $n = 2$ se determina ahora el número total de muestras que es posible obtener de esta población que tiene 5 elementos. La fórmula que se utiliza para determinar dicho número de muestras es la fórmula de las combinaciones que se vio en el capítulo 4 (en donde N es el número de elementos de la población y n es el número de elementos de la muestra):

$$C_n^N = \frac{N!}{n!(N - n)!} = \frac{5 \cdot 4 \cdot 3 \cdot 2}{2(3 \cdot 2)} = 10$$

Se tiene, entonces, que el número total de muestras de tamaño 2 que es posible obtener de una población con 5 elementos es de 10. En la tabla 7.4 aparece el listado de todas estas muestras, junto con sus correspondientes valores para la variable de número de hijos, la cual constituye la distribución muestral para $n = 2$.

Tabla 7.4 Distribución muestral para $n = 2$

Muestra	Hijos
Pérez, Gómez	2,4
Pérez, Durán	2,6
Pérez, Hidalgo	2,8
Pérez, Juárez	2,10
Gómez, Durán	4,6
Gómez, Hidalgo	4,8
Gómez, Juárez	4,10
Durán, Hidalgo	6,8
Durán, Juárez	6,10
Hidalgo, Juárez	8,10

En la tabla 7.5 se calcula la media de esas muestras (la distribución muestral de la media).

Tabla 7.5 Distribución muestral de la media

Muestra	Hijos	$\bar{X}$
Pérez, Gómez	2,4	3
Pérez, Durán	2,6	4
Pérez, Hidalgo	2,8	5
Pérez, Juárez	2,10	6
Gómez, Durán	4,6	5
Gómez, Hidalgo	4,8	6
Gómez, Juárez	4,10	7
Durán, Hidalgo	6,8	7
Durán, Juárez	6,10	8
Hidalgo, Juárez	8,10	9
Total		60

En la tabla 7.6, que contiene los mismos datos de la tabla 7.5, se calcula:

a) La media.
b) La desviación estándar de esta distribución muestral de la media.

Muestra	Hijos	$\bar{X}$	$(X - \bar{X})$	$(X - \bar{X})^2$
Pérez, Gómez	2,4	3	–3	9
Pérez, Durán	2,6	4	–2	4
Pérez, Hidalgo	2,8	5	–1	1
Pérez, Juárez	2,10	6	0	0
Gómez, Durán	4,6	5	–1	1
Gómez, Hidalgo	4,8	6	0	0
Gómez, Juárez	4,10	7	1	1
Durán, Hidalgo	6,8	7	1	1
Durán, Juárez	6,10	8	2	4
Hidalgo, Juárez	8,10	9	3	9
Total		60	0	30

a) La media de todas las medias, o sea la media de la distribución muestral de la media, a la que se representa mediante el símbolo $_{\bar{x}}$, es:

$$\mu_{\bar{x}} = \frac{\sum \bar{X}}{N} = \frac{60}{10}$$

La anterior es la conocida fórmula para determinar la media aritmética, sólo que en este caso se trata de la media de un conjunto de medias (y por ello el subíndice de μ) y N resulta ser el número de muestras. Además, como la distribución muestral de las medias contiene a *todas las muestras posibles*, entonces representa una población en sí y por eso la N es mayúscula. Además, si se piensa en términos probabilísticos, se dice que esta media de la distribución muestral de las medias es su valor esperado.

La primera conclusión de las anticipadas en el inciso 6 del procedimiento descrito se observa fácilmente en el resultado anterior, y consiste en que la media de la distribución muestral de las medias es igual a la media de la población o, dicho en otras palabras, el *valor esperado de la media es igual a la media de la población*. Esto mismo expresado en símbolos:

$$E(X) = \mu_{\bar{X}} = \frac{\sum \bar{X}}{n} \qquad \textbf{(7.1)}$$

b) La desviación estándar de todas las medias, o sea la desviación estándar de la distribución muestral de las medias, es:

$$\sigma_{\bar{x}} = \sqrt{\frac{\sum (\bar{X} - \mu_{\bar{x}})^2}{n}} = \sqrt{\frac{30}{10}} = \sqrt{3} = 1.732$$

En donde el subíndice $\bar{X}$ señala que es la desviación estándar de la distribución muestral de la media. A este valor se le conoce como *error estándar de la media* y tiene una relación con la desviación estándar de la población que se expresa mediante la siguiente ecuación:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N - n}{N - 1}} \qquad \textbf{(7.2)}$$

Para ilustrar que esta relación se cumple, se sustituyen los valores calculados (en el punto 2 anterior se determinó que la desviación estándar de la población, σ, es de 2.828):

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N - n}{N - 1}} = \frac{2.828}{\sqrt{2}} \sqrt{\frac{5 - 2}{5 - 1}}$$

$$= \frac{2.828}{1.414} \sqrt{0.75} = 2(0.866) = 1.73$$

que es el mismo valor que se encontró calculando la desviación estándar de la distribución muestral de la media, con el promedio de los cuadrados de las desviaciones entre cada una de las medias muestrales y la media de todas ellas. Y esta relación entre la desviación estándar de la población, σ y el error estándar de la media, o desviación estándar de la distribución muestral de la media es la segunda conclusión importante para el muestreo que se mencionó antes.

Antes de analizar con mayor detalle la importancia que las 2 conclusiones anteriores tienen para las técnicas de muestreo, conviene revisar la forma de la distribución muestral de las medias, ya que esto nos conducirá a una tercera conclusión, igualmente importante.

En la tabla 7.7 se presenta la distribución muestral de las medias, agrupadas según las frecuencias observadas.

Tabla 7.7 La distribución muestral de la media, agrupada según la frecuencia con que aparece cada valor

Media de la muestra	Frecuencia observada *f*
3	1
4	1
5	2
6	2
7	2

Media de la muestra	Frecuencia observada *f*
8	1
9	1
Total	10

Ahora, en la figura 7.1 se grafican los datos de la distribución de frecuencias de la tabla 7.5.

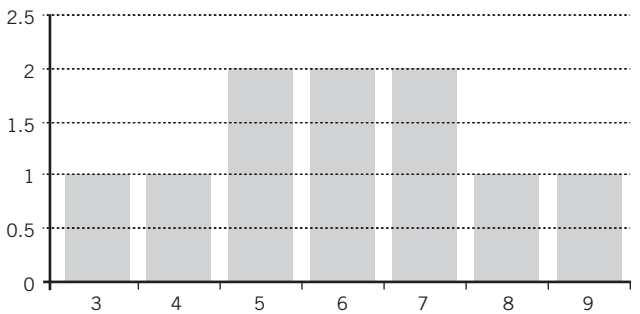


Figura 7.1 Gráfica de barras de la distribución muestral de la media.

En la gráfica anterior se aprecia que las medias de las muestras tienden a agruparse alrededor (cerca) del valor de la media de la población, lo cual es una característica de la distribución normal; aunque en este ejemplo el tamaño de la muestra es exageradamente pequeño, aun así se nota que la *distribución de las medias muestrales tiende a ser de forma normal*.

La afirmación anterior, expresada en el siguiente párrafo de manera más formal, es la última conclusión a la que se quería llegar y es lo que, junto con las dos primeras conclusiones revisadas antes, se conoce como el *teorema central del límite*.

Si X es una variable aleatoria para la que se conocen su media μ y su varianza σ^2, la distribución muestral de la media tiende a ser normal con media μ y desviación estándar (error estándar) $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$ conforme aumenta el tamaño de la muestra.

Es importante observar que el teorema central del límite no menciona la forma de la distribución de la población en la que se analiza la variable. Esto es de gran importancia porque, en términos prácticos, significa que la distribución muestral de la media tiende a ser normal aun cuando la variable no se distribuya de manera normal en la población, siempre y cuando la muestra sea de tamaño grande (que tienda a infinito). En la práctica, se considera que una muestra es grande cuando es igual o mayor que 30; como se verá más adelante, el tamaño de la muestra, junto con otras consideraciones, tiene efecto sobre el procedimiento específico que se debe seguir para estimar parámetros o para llevar a cabo pruebas de hipótesis.

A manera de recapitulación, y dada su importancia, se reproducen en seguida estas 3 conclusiones.

7.2.2 Tres conclusiones importantes que se desprenden de la distribución muestral de la media: el teorema central del límite

1. La media de la distribución muestral de las medias es igual a la media de la población o, dicho en otras palabras, el valor esperado de la media es igual a la media de la población. Esto mismo expresado en símbolos:

$$E(X) = \mu_{\bar{X}} = \frac{\sum \bar{X}}{n}$$

2. Existe una relación entre la desviación estándar de la población y la desviación estándar de la distribución muestral de las medias (a la que se conoce como *error estándar*) y que es:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

3. **Teorema central del límite.** Si X es una variable aleatoria para la que se conocen su media μ y su varianza σ^2, la distribución muestral de la media tiende a ser normal con media μ y desviación estándar (error estándar):

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

conforme aumenta el tamaño de la muestra.

7.2.3 Fórmula del error estándar de la media y factor de corrección por población finita

Para este análisis conviene observar el comportamiento de la parte $\sqrt{\frac{N-n}{N-1}}$ de la fórmula del error estándar. En primer lugar, si la población es infinita o muy numerosa, la fracción $\frac{N-n}{N-1}$ resulta ser prácticamente

igual a 1 y, por consiguiente, su raíz cuadrada también es igual a 1. Por ello, en los casos en los que se tiene una población muy cuantiosa o con una cantidad infinita de elementos, como este factor es igual a 1, multiplicarlo por la parte restante de la fórmula no tiene caso ya que se obtendrá el mismo resultado. En símbolos:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \frac{\sigma}{\sqrt{n}}(1) = \frac{\sigma}{\sqrt{n}}$$

Por ello, en estos casos, la fórmula del error estándar de la media se convierte en:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \tag{7.3}$$

Por otro lado, en los casos en los que el tamaño de la población es reducido en comparación con el tamaño de la muestra, es necesario evaluar la conveniencia de incluir este factor, al que se conoce como *factor de corrección por población finita* en los cálculos. Por ejemplo, si $N = 100$ y $n = 10$, el factor es:

$$\sqrt{\frac{N-n}{N-1}} = \sqrt{\frac{100-10}{100-1}} = \sqrt{\frac{90}{99}} = 0.9535$$

Si se omite este factor de corrección por población finita cuando se tienen estos tamaños de muestra y de población, se sobrestimaría el error estándar 5%. Si $N = 1\,000$ y $n = 30$, entonces

$$\sqrt{\frac{N-n}{N-1}} = \sqrt{\frac{1\,000-30}{1\,000-1}} = \sqrt{\frac{970}{999}} = 0.9854$$

y la sobrestimación que se acepta en el error estándar, si no se incluye este factor es ahora de sólo 1.5 por ciento.

En general (aunque esto depende en última instancia del estudio específico que se realiza) se dice que resulta aceptable eliminar este factor para simplificar la fórmula del error estándar, cuando la razón del tamaño de la muestra al tamaño de la población es menor o igual a 0.05. En símbolos, cuando,

$$\frac{n}{N} \leq 0.05$$

Por ejemplo, si $N = 600$ y $n = 30$:

$$\sqrt{\frac{N-n}{N-1}} = \sqrt{\frac{600-30}{600-1}} = \sqrt{\frac{570}{599}} = 0.9755$$

la eliminación del factor en este caso introduce una sobrestimación de sólo 2.5 por ciento.

Además, el hecho de que eliminar el factor produzca una *sobrestimación* del valor del error estándar, hace que su utilización resulte más conservadora, ya que se trabajaría con un error estándar (dispersión de la distribución muestral) mayor que la que se tendría si se incluye el factor.

7.2.4 Consideraciones adicionales sobre la distribución muestral de la media

Se revisan en seguida otros detalles sobre la distribución muestral analizada antes para ilustrar con mayor claridad por qué las distribuciones muestrales son el fundamento de la inferencia estadística.

Con todo lo anterior, se revisan en la siguiente sección algunos ejemplos prácticos.

EJEMPLO 7.2

Como se muestra, la distribución muestral de la media, los datos de la tabla 7.5 representan la totalidad de las muestras posibles, constituyen una población y, por ello, es posible construir su correspondiente distribución de probabilidad. Ésta se presenta en la tabla 7.8, junto con los demás datos.

Si se extrae de esa población de 5 familias *una sola muestra* de tamaño 2, ¿cuál es la probabilidad de que su media fuera de 6 hijos?

Tabla 7.8 Distribución de probabilidad de la distribución muestral de la media

Media de la muestra $\bar{X}$	Frecuencia observada f	Probabilidad de las medias muestrales $P(\bar{X})$
3	1	1/10 = 0.10
4	1	1/10 = 0.10
5	2	2/10 = 0.20
6	2	2/10 = 0.20
7	2	2/10 = 0.20
8	1	1/10 = 0.10
9	1	1/10 = 0.10
Total	10	

Solución: En esa tabla es fácil apreciar que existe una probabilidad de 20% de que la media de esa muestra fuese de 6. Por otro lado, la probabilidad de sacar una muestra cuya media esté entre 5 y 7 es de 60%. Esto ilustra que es altamente probable obtener una muestra que contenga el verdadero valor de la media de la población, al igual que la importancia de las 3 conclusiones a las que se llegó antes; en otras palabras *es altamente probable que una muestra extraída al azar de una población contenga el parámetro que se desea estimar, o con el cual se desean realizar pruebas de hipótesis.*

Ejemplo 7.3

Para ilustrar algunos otros puntos importantes se revisa ahora la distribución muestral de la media para muestras de tamaño 3.

1. Se utiliza la misma población de 5 familias.
2. $\mu = 6$, $\sigma = 2.828$, que ya se calculó antes.
3. $C_n^N = \dfrac{N!}{n!(N-n)!} = \dfrac{5!}{3!2!}\dfrac{5 \cdot 4 \cdot 3}{3 \cdot 2} = 10$

En la tabla 7.9 aparecen las muestras correspondientes, junto con los cálculos necesarios para la parte restante del ejemplo (los apellidos de las familias se identifican con su letra inicial).

Tabla 7.9 Datos para el ejemplo 7

Muestras	X Números de hijos	Media muestral $\bar{X}$	$(\bar{X} - \bar{x})$	$(\bar{X} - \bar{x})^2$
P,G,D	2,4,6	4	–2	4
P,G,H	2,4,8	4.67	–1.33	1.7689
P,G,J	2,4,10	5.33	–0.67	0.4489
P,D,H	2,6,8	5.33	–0.67	0.4489
P,D,J	2,6,10	6	0	0
P,H,J	2,8,10	6.67	0.67	0.4489
G,D,H	4,6,8	6	0	0
G,D,J	4,6,10	6.67	0.67	0.4489
G,H,J	4,8,10	7.33	1.33	1.7689
D,H,J	6,8,10	8	2	4
Total		60		13.3334

4. La distribución muestral de la media se muestra en la tercera columna de la tabla 7.9.
5. $\mu_{\bar{X}} = \dfrac{\sum \bar{X}}{n} = \dfrac{60}{10} = 6$

que es la misma media poblacional calculada antes.

Se calcula ahora la desviación estándar de esta distribución muestral (las operaciones aparecen en la tabla 7.9):

$$\sigma_{\bar{x}} = \sqrt{\frac{\sum(\bar{X} - \mu_{\bar{x}})^2}{n}} = \sqrt{\frac{13.3334}{10}} = \sqrt{1.3334} = 1.1547.$$

De acuerdo con la fórmula que relaciona el error estándar con la desviación estándar de la población:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \frac{2.828}{\sqrt{3}}\sqrt{\frac{5-3}{5-1}} = 1.6327\sqrt{0.5} = 1.1545$$

La ligera diferencia se debe a redondeo. Al igual que en el ejemplo 1, puede observarse en este otro ejemplo que la media de la distribución muestral de la media es igual a la media de la población.

$$E(X) = \mu_{\bar{X}} = \frac{\sum \bar{X}}{n}$$

Existe una relación entre la desviación estándar de la población y la desviación estándar de la distribución muestral de la media, que es:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}.$$

Además, comparando los resultados de los ejemplos 1 y 3, se puede deducir otra conclusión que es importante para el muestreo: *el error estándar disminuye al aumentar el tamaño de la muestra*. En el ejemplo 7.1 se encontró que el error estándar era de $\sigma_{\bar{x}} = 1.732$ con un tamaño de muestra $n = 2$, mientras que en el ejemplo 7.3 se determinó $\sigma_{\bar{x}} = 1.1545$ con un tamaño de muestra $n = 3$. Esta conclusión es importante porque, siendo el error estándar una medida de la dispersión de las medias de todas las muestras posibles, permite saber que, conforme mayor sea la muestra (una sola) que se extraiga para realizar un estudio muestral en la práctica, mayor será la precisión que se logre, puesto que menor será la dispersión de todas las muestras posibles.

Con todo lo anterior, se revisan en la siguiente sección algunos ejemplos prácticos.

7.2.5 Aplicaciones del análisis de la distribución muestral de la media

En esta parte se revisan algunos ejemplos en los que se ilustra cómo se utilizan las conclusiones extraídas del análisis de la distribución de la media para analizar situaciones reales. Vale la pena resaltar que es importante tener presente que se estará hablando de características de una población, de las de una muestra y, también, de las de la distribución muestral de la media y es de gran relevancia tener esto presente en cada caso.

EJEMPLO 7.4

Se extrae una muestra de $n = 30$ elementos de una población que se sabe que tiene un gran número de elementos y cuyas media y desviación estándar son $\mu = 162$ y $\sigma = 20$. Encuentre la probabilidad de que la media de esa muestra:

1. Sea superior a 170.
2. Esté entre 152 y 172.

Solución: Como se trata de extraer una sola muestra de entre todas las posibles (una cantidad prácticamente infinita de ellas), se mide la dispersión de estas medias muestrales mediante el error estándar. Además, como el tamaño de la población es prácticamente infinito, se puede utilizar la fórmula simplificada del error estándar (es decir, sin incluir el factor de corrección por población finita).

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{20}{\sqrt{30}} = \frac{20}{5.48} = 3.65$$

Como se trata de una muestra grande ($n = 30$), de acuerdo con el teorema central del límite, la distribución de las medias muestrales es aproximadamente normal, por lo que se utiliza el procedimiento de estandarizar los valores para convertirlos a unidades de la desviación estándar y así poder utilizar la tabla de áreas bajo la curva normal para determinar la probabilidad que se busca. Comenzando con el punto 1, la probabilidad de que la media de esa muestra sea superior a 170:

$$z = \frac{\bar{X} - \mu}{\sigma_{\bar{x}}} = \frac{170 - 162}{3.65} = 2.19$$

De la tabla de áreas bajo la curva normal:

$$P(z \geq 2.19) = 0.0143$$

por lo tanto, la probabilidad de que esa muestra tenga una media superior a 170 es de 1.43 por ciento.

2. La probabilidad de que la media de esa muestra esté entre 152 y 172:

$$z = \frac{152 - 162}{3.65} = -2.74$$

$$z = \frac{172 - 162}{3.65} = 2.74$$

De la tabla de áreas bajo la curva normal:

$$P(-2.74 \geq z \leq 2.74) = 0.4969 + 0.4969 = 0.9938$$

por lo que la probabilidad de que la media de esa muestra esté entre 152 y 172 es de 99.38%, es decir, está prácticamente asegurado que esté entre esos 2 valores.

EJEMPLO 7.5

El promedio del peso neto de atún enlatado por una empacadora es de $\mu = 325$ g, con una desviación estándar $\sigma = 20$ g, si se extrae de la producción de la empacadora una muestra aleatoria de 50 latas de atún, ¿cuál es la probabilidad de que su media:

1. sea inferior a 320 g?
2. esté entre 320 y 330 g?

Solución: La desviación estándar de la distribución muestral de la media (error estándar):

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{20}{\sqrt{50}} = \frac{20}{7.07} = 2.828$$

$$z = \frac{320 - 325}{2.828} = -1.77$$

1. De la tabla de áreas bajo la curva normal:

$$P(z \leq -1.77) = 0.5 - 0.4616 = 0.0384$$

por lo que la probabilidad de que la media de la muestra de 50 latas de atún tenga una media inferior a 320 g es de 3.84 por ciento.

2. Según se vio en punto 1, para una diferencia de 5 g respecto a la media, $z = 1.77$, por lo que

$$P(-1.77 \leq z \leq 1.77) = 0.4616 + 0.4616 = 0.9232.$$

En este ejemplo se puede apreciar la utilidad práctica del análisis de las distribuciones muestrales. Si efectivamente se saca la muestra y resulta que su media se encuentra entre 320 y 330 g, se concluiría que este resultado concuerda con los valores que se suponen para la población lo cual, a su vez, significaría que el proceso de empacado está bajo control; por otro lado, si la media de la muestra fuera, por ejemplo, de 300 g, se concluiría que ese resultado muestral no concuerda con las características que se suponen para la población (la hipótesis), ya que la probabilidad de extraer una muestra con esa media, dada una población como la supuesta, es prácticamente de 0. Por ello se sospecharía que el proceso está fuera de control y que sería necesario tomar medidas correctivas.

EJEMPLO 7.6

El promedio de los depósitos en cuentas de cheques de los 1 500 cuentahabientes de una sucursal bancaria es de \$4 000 con una desviación estándar de \$250. ¿Cuál es la probabilidad de seleccionar una muestra de 100 cuentahabientes que arroje una media muestral de menos de \$4 050?

Solución: Como se conocen los valores de n y N se utiliza el factor de corrección por población finita para calcular el error estándar:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \frac{250}{\sqrt{100}}\sqrt{\frac{1\,500-100}{1\,500-1}}$$

$$= 25(0.966414) = 24.16$$

$$z = \frac{4\,050 - 4\,000}{24.16} = 2.07$$

de la tabla de áreas bajo la curva normal,

$$P(z \leq 2.07) = 0.5 + 0.4808 = 0.9808$$

Así que la probabilidad de obtener una muestra de 100 cuentahabientes con media menor a \$4 050 es de 98.08 por ciento.

Aparte de la media aritmética, otra medida muy importante en los estudios por muestreo es la *proporción*. Se revisa en seguida, después de los ejercicios de esta sección, la distribución muestral de la proporción y se observa que el análisis conduce a conclusiones similares a las que se alcanzaron con la distribución muestral de la media.

EJERCICIOS 7.2 Distribución muestral de la media

1. Indique el total de muestras distintas que es posible obtener con los siguientes tamaños de muestras y de poblaciones:

a) $N = 100$; $n = 10$
b) $N = 1\,000$; $n = 2$
c) $N = 10$; $n = 5$
d) $N = \infty$; $n = 10$

2. En la siguiente tabla se encuentra el número de errores mecanográficos cometidos por 5 secretarias:

Secretaria	Núm. de errores
Alma	3
Susana	5
Gabriela	7
Corina	9
Marisa	11

Encuentre:

a) La media de esta población.
b) La desviación estándar de esta población.
c) La distribución muestral de las medias para muestras de tamaño $n = 2$.
d) El valor esperado de la media.
e) El error estándar de la media, mediante:
  i) Las medias de la distribución muestral de la media.
  ii) La desviación estándar de la población.
f) ¿Se cumple que
  i) $E(\overline{X}) = \mu$?
  ii) $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$?
g) Haga una gráfica de barras de la distribución muestral de la media, agrupando los valores según su frecuencia de aparición, ¿qué forma aproximada tiene la gráfica?
h) ¿Qué cantidad de medias muestrales se encuentran en el intervalo $\mu \pm 3\sigma_{\bar{x}}$?

3. Se tiene una población de 6 artículos y se determinó el número de defectos de fabricación de cada uno de ellos:

Artículo	Núm. de defectos
A	0
B	1
C	2
D	3
E	4
F	5

Encuentre:

a) La media de esta población.
b) La desviación estándar de esta población.

c) La distribución muestral de las medias para muestras de tamaño $n = 3$.
d) El valor esperado de la media.
e) El error estándar de la media, mediante:

i) Las medias de la distribución muestral de la media.
ii) La desviación estándar de la población.

f) ¿Se cumple que…

i) $E(\overline{X}) = \mu$?
ii) $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}\sqrt{\dfrac{N-n}{N-1}}$?

g) Haga una gráfica de barras de la distribución muestral de la media, agrupando los valores según su frecuencia de aparición, ¿qué forma aproximada tiene la gráfica?
h) ¿Qué cantidad de medias muestrales se encuentran en el intervalo $\mu \pm 3\sigma_{\bar{x}}$?

4. Se tiene una población con las siguientes características: $N = 1\,000$ y $\sigma = 40$. Encuentre el error estándar de la media para $n = 60$.
5. Se tiene una población con las siguientes características: $N = 1\,000$, y $\sigma = 40$. Encuentre el error estándar de la media para $n = 30$.
6. Se tiene una población muy numerosa con $\sigma = 40$. Encuentre el error estándar de la media para $n = 200$.
7. Se tiene una población con las siguientes características: $N = 1\,000$ y $\sigma = 40$. Encuentre el error estándar de la media para $n = 100$.
8. ¿Qué sucede con el error estándar cuando se aumenta el tamaño de la muestra?
9. Se tiene una población con $N = 1\,000$, $\mu = 100$ y $\sigma = 10$. Calcule la probabilidad de extraer de esta población una muestra de $n = 30$ elementos que tenga una media:

a) Inferior a 95.
b) Superior a 102.
c) Entre 97 y 103.

10. Se tiene una población con $N = 10\,000$, $\mu = 20$ y $\sigma = 2$. Calcule la probabilidad de extraer de esta población una muestra de $n = 100$ elementos que tenga una media:

a) Inferior a 19.5.
b) Superior a 20.1.
c) Entre 19.7 y 20.4.

11. De acuerdo con los registros históricos de un rastro, el peso promedio de las reses que llegan al matadero es de 750 kg con desviación estándar de 150 kg. Determine la probabilidad de que el peso promedio de una muestra de $n = 100$ de esas reses esté:

a) Por encima de 780 kg.
b) Por debajo de 735 kg.
c) Entre 740 y 760 kg.

7.3 Distribución muestral de la proporción

Se vio en las secciones anteriores que:

- La distribución muestral es el conjunto de todas las muestras de tamaño n que se pueden sacar de una población de tamaño N.
- Se puede calcular el número de elementos de una distribución muestral así definida como el número de combinaciones de N elementos tomados de n en n:

$$C_n^N = \frac{N!}{n!(N-n)!}.$$

- Si se calculan las medias de todas las muestras de una distribución muestral, entonces se tiene la distribución muestral de las medias.
- Desprendido directamente de lo anterior, la distribución muestral de la proporción es el conjunto de las proporciones de todas las muestras de tamaño n que se pueden sacar de una población de tamaño N.

En esta sección se analiza este tema con mayor detalle.

7.3.1 Desarrollo

En la tabla siguiente se comparan las definiciones de las distribuciones muestral de la media y de la proporción.

Distribución muestral de la media	Distribución muestral de la proporción
El conjunto de las medias de todas las muestras posibles de determinado tamaño n que es posible obtener de una determinada población de tamaño N.	El conjunto de las proporciones de todas las muestras posibles de determinado tamaño n que es posible obtener de una determinada población de tamaño N.

El procedimiento que se sigue para ilustrar la distribución muestral de la proporción es el mismo que se utilizó para la distribución muestral de la media:

1. Se supondrá una población con N elementos.
2. Se calculará la proporción y la desviación estándar de la población.
3. Se determinará el número total de muestras distintas de tamaño n que es posible extraer y se encontrará cuáles son; es decir, se enumerará la distribución muestral (el conjunto de todas las muestras de ese tamaño que es posible obtener de esa población).
4. Se decidirá la proporción de cada una de las muestras identificadas en el paso 3, con esto ya se tiene la distribución muestral de la proporción.
5. Se calcularán la media y la desviación estándar de la distribución muestral de la proporción.
6. Se obtendrán conclusiones similares a las conseguidas para la distribución muestral de la media, y que son igualmente importantes para el muestreo, al comparar los parámetros poblacionales con los valores obtenidos para la distribución muestral de la proporción.
7. Se observará también la *forma* de la distribución muestral de la proporción para revisar la otra importante conclusión (la que se refiere a la normalidad de la distribución muestral cuando el tamaño de la muestra es grande).

Ejemplo 7.7

Suponga una población de $N = 6$ artículos, de los cuales 3 están defectuosos y 3 no. Si se utiliza "1" para representar la característica de estar defectuoso y "0" para representar la característica de no estarlo, los datos de la población se presentan en la tabla 7.10. En esta tabla se incluyen los cálculos necesarios para determinar la proporción y la desviación estándar de esta población.

Tabla 7.10 Datos para el ejemplo 7.7

Artículo	X	$(x - \pi)$	$(X - \pi)^2$
A	1	0.5	0.25
B	1	0.5	0.25
C	1	0.5	0.25
D	0	–0.5	0.25
E	0	–0.5	0.25
F	0	–0.5	0.25
Total	3	0	1.50

2. La proporción de la población:

$$\pi = \frac{\sum X}{N} = \frac{3}{6} = 0.5$$

La desviación estándar de la población:

$$\sigma = \sqrt{\frac{\sum (X - \pi)^2}{N}} = \sqrt{\frac{1.5}{6}} = \sqrt{0.25} = 0.5$$

La cual también puede calcularse como:

$$\sigma = \sqrt{\pi Q} = \sqrt{0.5(0.5)} = 0.5$$

3. El número total de muestras distintas de tamaño $n = 2$ que es posible obtener:

$$C_n^N = \frac{N!}{n!(N - n)!} = \frac{6!}{2!4!} = \frac{6 \cdot 5}{2} = 15$$

4. En la tabla 7.11 se presenta esta distribución muestral, junto con sus correspondientes proporciones.

Tabla 7.11 Distribución muestral de la proporción para muestras de tamaño $n = 2$

Muestras	X	Proporción (p)
a,b	1,1	2/2 = 1
a,c	1,1	2/2 = 1
a,d	1,0	1/2 = 0.5
a,e	1,0	1/2 = 0.5
a,f	1,0	1/2 = 0.5
b,c	1,1	2/2 = 1
b,d	1,0	1/2 = 0.5
b,e	1,0	1/2 = 0.5
b,f	1,0	1/2 = 0.5
c,d	1,0	1/2 = 0.5
c,e	1,0	1/2 = 0.5
c,f	1,0	1/2 = 0.5
d,e	0,0	0/2 = 0
d,f	0,0	0/2 = 0
e,f	0,0	0/2 = 0
Total	0,0	7.5

5. La media de esta distribución muestral de proporciones es:

$$\mu_p = \frac{\sum p}{N} = \frac{7.5}{15} = 0.5$$

que es igual a la proporción de la población. De manera similar a la conclusión de que la media de la distribución muestral de medias es igual a la media de la población, la de la proporción es igual a la proporción de la población.

Al igual que en el caso de la distribución muestral de la media, como la distribución muestral de la proporción es una

población, en el sentido de que se trata de las proporciones de todas las muestras posibles, su media es, al mismo tiempo, lo que se conoce como *valor esperado*, $E(p)$. Resumiendo esto en símbolos:

$$E(p) = \mu_p = \pi \tag{7.4}$$

Para determinar la desviación estándar de esta distribución muestral de proporciones, en la tabla 7.12 se presenta la misma distribución, pero agrupada de acuerdo con la frecuencia de las proporciones muestrales. En la tabla se incluyen las operaciones necesarias para calcular la desviación estándar.

Tabla 7.12 Distribución muestral de proporciones agrupadas de acuerdo con su frecuencia

Proporciones muestrales p	Frecuencias f	$(p - \mu_p)$	$(p - \mu_p)^2$	$f(p - \mu_p)^2$
0	3	– 0.5	0.25	0.75
0.5	9	0	0	0
1	3	0.5	0.25	0.75
Total	15	0		1.5

$$\sigma_p = \sqrt{\frac{\sum (p - \mu_p)^2}{n}} = \sqrt{\frac{1.5}{15}} = \sqrt{0.1} = 0.3162$$

De nueva cuenta, y al igual que sucede con la distribución muestral de la media, existe una relación entre esta desviación estándar de la distribución muestral de la proporción, σ_p, y la desviación estándar de la población, p:

$$\sigma_p = \sqrt{\frac{\pi Q}{n}} \sqrt{\frac{N - n}{N - 1}} \tag{7.5}$$

Para verificar que se cumple con los datos del ejemplo, se sustituyen los valores hallados antes:

$$\sigma_p = \sqrt{\frac{\pi Q}{n}} \sqrt{\frac{N - n}{N - 1}} = \frac{0.5}{\sqrt{2}} \sqrt{\frac{6 - 2}{6 - 1}}$$

$$= 0.353553(0.894427) = 0.3162$$

que es el mismo valor encontrado a partir de la distribución muestral.

6. En el paso anterior se compararon los valores de la población con los de la distribución muestral para llegar a las conclusiones apuntadas, las cuales, como se señaló, son paralelas a las que se obtuvieron para la distribución muestral de la media.
7. Finalmente, observando la distribución de frecuencias de la tabla 7.12, es fácil apreciar que, aun para una muestra tan pequeña como $n = 2$, la distribución de las proporciones muestrales tiende a la normalidad (9 observaciones en el centro y 3 en cada extremo), lo cual también coincide con la tendencia observada en la distribución muestral de la media.

Como es importante tanto para las aplicaciones del muestreo como para la comprensión por parte de los estudiantes, en las secciones siguientes se reproducen para las proporciones las secciones en que se dividió el tema de la distribución muestral de la media:

- Tres conclusiones importantes sobre la distribución muestral de la proporción.
- El factor de corrección por población finita.
- Consideraciones adicionales sobre la distribución muestral de la proporción.

7.3.2 Tres conclusiones importantes sobre la distribución muestral de la proporción

Las 3 conclusiones que se resumieron en la sección 7.2.2 son paralelas a éstas, sólo que aquí se trata de proporciones y en aquel caso se trataba de medias.

1. La media de la distribución muestral de proporciones (el valor esperado de la proporción) es igual a la proporción de la población:

$$E(p) = \mu_p = \pi$$

2. Existe una relación entre la desviación estándar de la población (binomial) y la desviación estándar de la distribución muestral de la proporción (error estándar):

$$\sigma_p = \sqrt{\frac{\pi Q}{n}} \sqrt{\frac{N - n}{N - 1}}$$

3. **Teorema central del límite.** Si X es una variable aleatoria para la que se conocen su proporción π y su varianza σ^2, la distribución muestral de la proporción tiende a ser normal con media μ_p y desviación estándar (error estándar):

$$\sigma_p = \sqrt{\frac{\pi Q}{n}}\sqrt{\frac{N-n}{N-1}}$$

conforme aumenta el tamaño de la muestra.

La única diferencia en la interpretación de este teorema consiste en observar que la población a la que se refiere es una población binomial y, por lo tanto, $\mu = \pi$, lo cual, en otras palabras, quiere decir que la media de una población binomial es su proporción.

7.3.3 Fórmula del error estándar de la proporción y factor de corrección por población finita

Al igual que antes, es posible eliminar el factor de corrección por población finita de la fórmula del error estándar cuando $\frac{n}{N} \leq 0.05$ o cuando la población tiene una cantidad muy grande o infinita de elementos. Así, la forma simplificada del error estándar en estos casos sería:

$$\sigma_p = \frac{\sigma}{\sqrt{n}} = \frac{\sqrt{\pi Q}}{\sqrt{n}} = \sqrt{\frac{\pi Q}{n}} \tag{7.6}$$

EJEMPLO 7.8

Una cadena de tiendas de departamentos tiene 5 000 cuentas de crédito abiertas con sus clientes. Se sabe, de acuerdo con la experiencia de varios años que la proporción de cuentas de crédito que se encuentran atrasadas en sus pagos (morosas) es de 10%; si se extrae de esta población una muestra aleatoria de 100 cuentas, ¿cuál es la probabilidad de que más de 13 de ellas sean morosas?

Solución: La desviación estándar de la población:

$$\sigma = \sqrt{\pi Q} = \sqrt{0.1(0.9)} = 0.3$$

Como $$\frac{n}{N} = \frac{100}{5\,000} = 0.02 < 0.05$$

no se utiliza el factor de corrección por población finita y se aplica la forma simplificada de la fórmula del error estándar (desviación estándar de la distribución muestral) de la proporción,

$$\sigma_p = \frac{\sigma}{\sqrt{n}} = \frac{0.3}{\sqrt{100}} = 0.03$$

Además, como se trata de una muestra grande ($n > 30$), se puede utilizar el teorema central del límite y considerar que la distribución del conjunto de todas las muestras implicadas tiene una forma aproximadamente normal. La proporción de la muestra:

$$p = \frac{X}{n} = \frac{13}{100} = 0.13$$

y

$$z = \frac{p - \pi}{\sigma_p} = \frac{0.13 - 0.10}{0.03} = \frac{0.03}{0.03} = 1$$

De la tabla de áreas bajo la curva normal:

$$P(z \geq 1) = 0.5 - 0.3413 = 0.1587$$

por lo que la probabilidad de que en esa muestra haya más de 13 cuentas morosas es de 15.87 por ciento.

EJEMPLO 7.9

Una empresa cuenta con 1 000 artículos en su inventario, de acuerdo con la experiencia que se tiene con las auditorías del inventario, la proporción de registros contables por artículo que no coinciden con el inventario que realmente se tiene disponible (el inventario físico) es de 15%; si se extrae una muestra de 100 registros contables, ¿cuál es la probabilidad de que la proporción muestral de registros que no coinciden con el inventario físico sea mayor de 20 por ciento.

Solución: En este caso, como

$$\frac{n}{N} = \frac{100}{1\,000} = 0.1 > 0.05$$

sí es necesario incluir el factor de corrección por población finita en la fórmula del error estándar, pero antes de calcular este error estándar, conviene revisar alguna otra forma alternativa de la fórmula completa (son simples manipulaciones algebraicas):

$$\sigma_p = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \frac{\sqrt{\pi Q}}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

$$= \sqrt{\frac{\pi Q}{n}}\sqrt{\frac{N-n}{N-1}} = \sqrt{\frac{\pi Q}{n}}\cdot\frac{N-n}{N-1}.$$

Por supuesto, el lector puede utilizar la forma que más conveniente le resulte, aunque en las operaciones que siguen se utiliza la forma en la que aparece por separado el factor de corrección por población finita para realzar el efecto que su inclusión tiene sobre el valor del error estándar

$$\sigma_p = \sqrt{\frac{\pi Q}{n}}\sqrt{\frac{N-n}{N-1}} = \sqrt{\frac{0.15(0.86)}{100}}\sqrt{\frac{1\,000-100}{1\,000-1}}$$
$$= 0.0357(0.949) = 0.0339$$

Como de nueva cuenta se trata de una muestra grande ($n > 30$) se puede aplicar el teorema central del límite para utilizar la distribución normal en el cálculo de la probabilidad:

$$z = \frac{p-\pi}{\sigma_p} = \frac{0.20-0.15}{0.0339} = 1.47$$

de la tabla de áreas bajo la curva normal:

$$P(z \geq 1.47) = 0.5 - 0.4292 = 0.0708$$

por lo que la probabilidad de que esa muestra contenga más de 20 registros contables que no coincidan con el inventario físico es de 7.08 por ciento.

7.3.4 Consideraciones adicionales sobre la distribución muestral de la proporción

Aquí, al igual que sucede con la media aritmética, el análisis de la distribución muestral de la proporción permite concluir que:

- Es altamente probable que una muestra extraída al azar de una población contenga el parámetro, la proporción que se desea estimar, o con el cual se desean realizar pruebas de hipótesis.
- El error estándar disminuye al aumentar el tamaño de la muestra; esta conclusión se puede deducir de un análisis algebraico de la fórmula simplificada del error estándar. Para encontrar este error estándar se divide la desviación estándar de la población entre la raíz cuadrada del tamaño de la muestra, al aumentar éste, aumenta el valor de dicha raíz cuadrada y por lo tanto disminuye el cociente.

EJEMPLO 7.10

Encuentre el error estándar para una situación en la que la $\pi = 0.7$ cuando:

1. $n = 10$.
2. $n = 100$.

Solución:

1. $\sigma_p = \sqrt{\frac{\pi Q}{n}} = \sqrt{\frac{0.7(0.3)}{10}} = 1.45$

2. $\sigma_p = \sqrt{\frac{\pi Q}{n}} = \sqrt{\frac{0.7(0.3)}{100}} = 0.046$

lo cual ilustra lo apuntado antes, que el error estándar disminuye cuando aumenta el tamaño de la muestra.

EJERCICIOS 7.3 Distribución muestral de la proporción

1. Considere que se tiene una población de 5 personas, de las cuales 2 son casadas y las otras 3 son solteras. Desde el punto de vista de la proporción de casados, encuentre:

a) La proporción de esta población.
b) La desviación estándar de esta población.
c) La distribución muestral de las proporciones para muestras de tamaño $n = 2$.
d) El valor esperado de la proporción.
e) El error estándar de la proporción, mediante:
  i) las proporciones de la distribución muestral de la proporción
  ii) la desviación estándar de la población
f) ¿Se cumple que
  i) $E(p) = \mu_p = \pi$?
  ii) $\sigma_p = \sqrt{\frac{\pi Q}{n}}\sqrt{\frac{N-n}{N-1}}$?

g) Haga una gráfica de barras de la distribución muestral de la proporción agrupando los valores según su frecuencia de aparición, ¿qué forma aproximada tiene la gráfica?
h) ¿Qué cantidad de proporciones muestrales se encuentran en el intervalo $\pi \pm 2\sigma_p$?

2. Se tiene una población de 6 niños de 6 años de edad y se determinó que 3 de ellos saben leer con eficiencia. Desde el punto de vista de los que saben leer con eficiencia, encuentre:

a) La proporción de esta población.
b) La desviación estándar de esta población.
c) La distribución muestral de las proporciones para muestras de tamaño $n = 4$.
d) El valor esperado de la proporción.
e) El error estándar de la proporción, mediante:

i) las proporciones de la distribución muestral de la proporción
ii) la desviación estándar de la población

f) ¿Se cumple que

i) $E(p) = \mu_p = \pi$?

ii) $\sigma_p = \sqrt{\dfrac{\pi Q}{n}}\sqrt{\dfrac{N-n}{N-1}}$?

g) Haga una gráfica de barras de la distribución muestral de la proporción, agrupando los valores según su frecuencia de aparición, ¿qué forma aproximada tiene la gráfica?
h) ¿Qué cantidad de medias muestrales se encuentran en el intervalo $\pi \pm 3\sigma_p$?

3. Se tiene una población con las siguientes características: $N = 1\ 000$, $n = 60$ y $p = 0.7$. Encuentre el error estándar de la media.

4. Se tiene una población con las siguientes características: $N = 1\ 000$, $n = 30$ y $s = 0.7$. Encuentre el error estándar de la proporción.

5. Se tiene una población muy numerosa y con las siguientes características: $n = 200$ y $p = 0.6$. Encuentre el error estándar de la proporción.

6. Se tiene una población con las siguientes características: $N = 1\ 000$, $n = 100$ y $p = 0.4$. Encuentre el error estándar de la proporción.

7. ¿Qué sucede con el error estándar de la proporción cuando disminuye el tamaño de la muestra?

8. Se tiene una población con $N = 1\ 000$ y $p = 0.35$. Calcule la probabilidad de extraer de esta población una muestra de $n = 30$ elementos que tenga una proporción:

a) inferior a 0.20.
b) superior a 0.40.
c) entre 0.25 y 0.45.

9. Se tiene una población con $N = 10\ 000$ y $p = 0.10$. Calcule la probabilidad de extraer de esta población un muestra de $n = 100$ elementos que tenga una proporción:

a) inferior a 0.09.
b) superior a 0.15.
c) entre 0.08 y 0.17.

10. De acuerdo con los registros históricos de un banco, la proporción de clientes que tienen tarjeta de crédito y que no se atrasan en sus pagos es de 70%. Determine la probabilidad de que, en una muestra de 50 de esos clientes, la proporción de los que están al corriente en sus pagos esté:

a) por encima de 0.80.
b) por debajo de 0.65.
c) entre 0.73 y 0.80.

7.4 Distribución muestral de la varianza

El propósito fundamental que se persigue al analizar la distribución muestral de la varianza es ilustrar una característica a la que se denomina *sesgo*. En las 2 distribuciones muestrales anteriores (de la media y de la proporción) se llegó a la conclusión de que la media de la correspondiente distribución muestral era igual al parámetro de la población.

Se vio también que cuando se extrae una muestra y se utiliza el correspondiente estadístico muestral para la estimación de su parámetro correspondiente, se dice que el estadístico se convierte en *estimador*. Así, cuando se utiliza la media de una muestra para estimar la media de la población, se usa el estadístico muestral como estimador; de la misma manera, se emplea como estimador de la proporción poblacional a la proporción de la muestra. En muchos estudios por muestreo no se conoce la varianza de la población y, por ello, es necesario estimarla con base en la varianza de la muestra, y es por esto último que se vuelve importante revisar la distribución muestral de la varianza con el propósito de averiguar la forma en la que se hará la estimación. Para enfatizar este importante punto:

- La media de la distribución muestral de la media es igual a la media de la población.
- La media de la distribución muestral de la proporción es igual a la proporción de la población.

Estas características son las que permiten definir la media y la proporción como estimadores insesgados. Por otro lado, como no se cumple que la media de la distribución muestral de la varianza es igual a la varianza de la población, se dice que la varianza es un *estimador sesgado*.

Ejemplo 7.11

Se utilizan de nuevo los datos que se vieron en el ejemplo 7.1. La población consta de 5 familias y se determinó el número de hijos de cada una de ellas, posteriormente se obtuvo la distribución muestral para $n = 2$ y se construyó la distribución muestral de la media. Para mejor referencia se reproducen en seguida los datos de la población en la tabla 7.13, junto con los cálculos y los valores de su media y su desviación estándar.

Tabla 7.13 Número de hijos

Familia	Hijos X	$(X - \mu)$	$(X - \mu)^2$
Pérez	2	–4	16
Gómez	4	–2	4
Durán	6	0	0
Hidalgo	8	2	4
Juárez	10	4	16
Totales	30	0	40

La media aritmética de esta población.

$$\mu = \frac{\sum X}{N} = \frac{30}{5} = 6$$

La desviación estándar:

$$\sigma = \sqrt{\frac{\sum (X - \mu)^2}{N}} = \sqrt{\frac{40}{5}} = \sqrt{8} = 2.8284271$$

De donde, la varianza:

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N} = \frac{40}{5} = 8\,.$$

Ahora se va a construir la distribución muestral de la varianza calculando la varianza de cada una de las muestras de tamaño 2. En la tabla 7.14 se presentan estos datos.

Tabla 7.14 La distribución muestral de la varianza

Muestra	Hijos	$\bar{X}$	Varianzas muestrales s^2	Varianzas modificadas
Pérez, Gómez	2,4	3	1	2
Pérez, Durán	2,6	4	4	8
Pérez, Hidalgo	2,8	5	9	18
Pérez, Juárez	2,10	6	16	32
Gómez, Durán	4,6	5	1	2
Gómez, Hidalgo	4,8	6	4	8
Gómez, Juárez	4,10	7	9	18
Durán, Hidalgo	6,8	7	1	2
Durán, Juárez	6,10	8	4	8
Hidalgo, Juárez	8,10	9	1	2
Total		60	50	100

La columna de las varianzas (la distribución muestral de la varianza) se calculó de la siguiente manera tomando por ejemplo la varianza de la primera muestra, la correspondiente a las familias Pérez y Gómez y que es igual a 1:

$$s^2 = \frac{\sum (X - \bar{X})^2}{n} = \frac{(2-3)^2 + (4-3)^2}{2}$$

$$= \frac{(-1)^2 + 1^2}{2} = \frac{2}{2} = 1$$

La varianza de la cuarta muestra (16):

$$s^2 = \frac{(2-6)^2 + (10-6)^2}{2} = \frac{-4^2 + 4^2}{2} = \frac{16+16}{2} = 16$$

las demás varianzas fueron calculadas de manera similar.

Se acaba de reproducir, en párrafos anteriores, que en el ejemplo 7.1 se encontró $\sigma = 2.828$, por lo que la varianza de la población es

$$\sigma^2 = 2.828^2 = 8$$

Ahora, utilizando los datos de la tabla 7.14, se calcula la media de la distribución muestral de la varianza:

$$\mu_{s^2} = \frac{\sum S_i}{n} = \frac{50}{10} = 5$$

se observa claramente que, a diferencia de los otros 2 casos, la media de la distribución muestral de la varianza ***no*** es igual a la varianza de la población:

$$_{s^2} = 5 \neq 8 = \sigma^2.$$

Por esta característica se dice que la varianza muestral es un *estimador sesgado* de la varianza de la población, mientras que la proporción y la media son *estimadores insesgados* de sus correspondientes parámetros.

Cuando se calcula la varianza se divide a la suma de los cuadrados de las diferencias entre N, cuando se trata de la varianza de la población, y entre n cuando se trata de la varianza de la muestra.

Para utilizar la varianza muestral como estimador de la varianza de la población se utiliza una forma modificada de la varianza. La modificación consiste en dividir la suma de los cuadrados de las diferencias entre $N - 1$, en vez de dividirla entre N o, equivalentemente, dividirla entre $n - 1$, en vez de dividir entre n, según corresponda al caso de una población o de una muestra, respectivamente.

Así, cuando la fórmula de la varianza de una población es:

$$\sigma^2 = \frac{\sum(X - \mu)^2}{N}.$$

La varianza modificada de la población es:

$$\sigma^2 = \frac{\sum(X - \mu)^2}{N - 1} \qquad \textbf{(7.7)}$$

Se tiene entonces que la varianza modificada de la población es:

$$\sigma^2 = \frac{\sum(X - \mu)^2}{N - 1} = \frac{40}{4} = 10.$$

Consecuentemente, la varianza modificada de cada una de las muestras es:

$$s^2 = \frac{\sum(X - \mu)^2}{n - 1} \qquad \textbf{(7.8)}$$

En la última columna de la tabla 7.14 aparecen las varianzas muestrales modificadas. Por ejemplo, la varianza modificada de la primera muestra se calculó:

$$S^2 = \frac{\sum(X - \bar{X})^2}{n - 1} = \frac{2}{1} = 2$$

La varianza modificada de la cuarta muestra a su vez:

$$S^2 = \frac{\sum(X - \bar{X})^2}{n - 1} = \frac{32}{1} = 32$$

Entonces resulta que el conjunto de valores de la última columna constituyen la distribución muestral de la varianza modificada. Calculando ahora la media de esta distribución muestral:

$$\mu_{s_m^2} = \frac{\sum s_m^2}{n} = \frac{100}{10} = 10$$

Se puede apreciar la conclusión a la que se quería llegar, que la media de la distribución muestral de la varianza modificada sí es igual a la varianza modificada de la población o, en otras palabras, la varianza muestral modificada es un estimador insesgado de la varianza poblacional modificada. Se resumen los resultados anteriores en la tabla 7.15.

Tabla 7.15 Resumen de los resultados de la distribución muestral de la varianza

Varianza de la población, σ^2	8
Media de la distribución muestral de la varianza para $n = 2$	5
Varianza modificada de la población	10
Media de la distribución muestral de la varianza modificada para $n = 2$	10

Como puede verse en estos datos, la media de la distribución muestral de la varianza, 5, subestima en 3 unidades a la de la población, 8, o 3/8 = 0.375 (37.5%), en tanto que la correspondiente a la varianza modificada, 10, la sobrestima, pero sólo en 2 unidades, que son 2/8 = 0.25, o sea 25 por ciento.

7.4.1 Distribuciones muestrales sin reemplazo y con reemplazo

Todas las distribuciones muestrales construidas hasta aquí concuerdan con el muestreo en la práctica, se hacen sin reemplazo y su número se determina con la fórmula de las combinaciones como ya vimos y que es:

$$C_n^N = \frac{N!}{n!(N - n)!}$$

Sin embargo, desde el punto de vista teórico, también se puede considerar "el conjunto de todas las muestras posibles", como se definió a las distribuciones muestrales, considerando que el muestreo se hace *con reemplazo*. En este caso, el número de muestras posibles es considerablemente mayor, en el ejemplo que se sigue, el del número de hijos de 5 familias y los valores de las variables eran 2, 4, 6, 8 y 10. Con esta misma población se pueden construir 25 muestras, si se acepta el reemplazo de los elementos. En el ejemplo siguiente se muestra lo que sucede con la distribución muestral de la varianza (DMV) cuando el muestreo se hace con reemplazo. En la tabla 7.16 tenemos el conjunto de esas 25 muestras que conforman ahora la DMV, junto con los cálculos de medias, varianzas y la media de la DMV.

Tabla 7.16 La DMV cuando se permite el reemplazo

Muestra	Elemento 1	Elemento 2	Varianzas muestrales s^2	Varianzas modificadas s_m^2
1	2	4	1	2
2	2	6	4	8
3	2	8	9	18
4	2	10	16	32

(*continúa*)

Tabla 7.16 (*continuación*)

Muestra	Elemento 1	Elemento 2	Varianzas muestrales s^2	Varianzas modificadas s_m^2
5	4	6	1	2
6	4	8	4	8
7	4	10	9	18
8	6	8	1	2
9	6	10	4	8
10	8	10	1	2
11	2	2	0	0
12	4	4	0	0
13	6	6	0	0
14	8	8	0	0
15	10	10	0	0
16	4	2	1	2
17	6	2	4	8
18	8	2	9	18
19	10	2	16	32
20	6	4	1	2
21	8	4	4	8
22	10	4	9	18
23	8	6	1	2
24	10	6	4	8
25	10	8	1	2
		Promedios	4	8

En la tabla anterior puede verse que la media de la distribución muestral de la varianza en muestreo con reemplazo está todavía más alejada de la media de la población, en tanto que la media de esta misma distribución muestral de la varianza modificada con reemplazo es exactamente igual a la varianza poblacional.

Esto, junto con los resultados que se obtuvieron antes respecto a esta DMV con muestreo sin reemplazo es lo que conduce a concluir que el mejor estimador de la varianza poblacional es, precisamente, la varianza modificada.

En la tabla 7.17 se reproducen los resultados ya obtenidos y que se presentaron en la tabla 7.15, y a los que se añaden los que se acaban de obtener.

Tabla 7.17 Resumen de los resultados de la DMV en muestreo sin y con reemplazo

Muestreo sin reemplazo	
Varianza de la población, σ^2	8
Media de la distribución muestral de la varianza para $n = 2$	5
Varianza modificada de la población	10
Media de la distribución muestral de la varianza modificada para $n = 2$	10
Muestreo con reemplazo	
Media de la distribución muestral de la varianza para $n = 2$	4
Media de la distribución muestral de la varianza modificada para $n = 2$	8

En lo sucesivo, cuando se utilice la varianza muestral para estimar la varianza de la población se emplea la varianza modificada y ya no se utiliza el subíndice μ que se usó aquí para distinguirla de la varianza común:

$$s^2 = \frac{\sum(X - \bar{X})^2}{n - 1}$$

Al mismo tiempo, como la desviación estándar es la raíz cuadrada de la varianza, al usar la desviación estándar de la muestra, s, como estimador de la desviación estándar de la población, σ, se realizarán los cálculos con $n - 1$ en el denominador:

$$s = \sqrt{\frac{\sum(X - \bar{X})^2}{n - 1}}$$

7.4.2 Estimadores insesgados y estimadores sesgados

Dada su importancia, se resumen aquí las conclusiones planteadas en la sección anterior.

- La media es un estimador insesgado porque la media de su distribución muestral (su valor esperado) es igual a la media de la población:

$$E(\bar{X}) = \mu_X = \mu$$

- La proporción es un estimador insesgado porque la media de su distribución muestral (su valor esperado) es igual a la proporción de la población:

$$E(P) = \mu_p = \pi$$

- La varianza es un estimador sesgado porque la media de su distribución muestral (su valor esperado) no es igual a la varianza de la población:

$$E(S^2) = \mu_{S^2} \neq \sigma^2$$

- La varianza modificada es un estimador insesgado porque la media de su distribución muestral (su valor esperado) es igual a la varianza modificada de la población:

$$E(S_m^2) = \mu_{S_m^2} = \sigma_m^2$$

ejercicios 7.4 Distribución muestral de la varianza y estimadores sesgados e insesgados

1. En la siguiente tabla se encuentra el número de errores mecanográficos cometidos por 5 secretarias.

Secretaria	Núm. de errores
Alma	3
Susana	5
Gabriela	7
Corina	9
Marisa	11

Como puede ver, son los mismos datos que se presentan en la actividad 5 de los ejercicios 7.1. Determine:

a) La varianza de la población.
b) La varianza modificada de la población.
c) La distribución muestral de la varianza para $n = 2$.
d) La distribución muestral de la varianza modificada para $n = 2$.
e) El valor esperado de la varianza.
f) El valor esperado de la varianza modificada.
g) ¿Se cumple que $E(S^2) = \mu_{S^2} \neq \sigma^2$?
h) ¿Se cumple que $(E(S_m^2) = \mu_{S_m^2} = \sigma_m^2)$?

2. Se tiene una población de 6 artículos y se determinó el número de defectos de fabricación de cada uno de ellos:

Artículo	Núm. de defectos
A	0
B	1
C	2
D	3
E	4
F	5

Como puede ver, se trata de los mismos datos que se presentan en la actividad 6 de los ejercicios 7.1. Determine:

a) La varianza de la población.
b) La varianza modificada de la población.
c) La distribución muestral de la varianza para $n = 3$.
d) La distribución muestral de la varianza modificada para $n = 3$.
e) El valor esperado de la varianza.
f) El valor esperado de la varianza modificada.
g) ¿Se cumple que $(E(S^2) = \mu_{S^2} \neq \sigma^2)$?
h) ¿Se cumple que $(E(S_m^2) = \mu_{S_m^2} = \sigma_m^2)$?

3. ¿Por qué se dice que la varianza muestral es un *estimador sesgado de la varianza de la población*?
4. ¿Por qué se dice que la varianza muestral modificada es un *estimador insesgado de la varianza modificada de la población*?

7.5 Resumen

Las conclusiones que se pueden ilustrar revisando las distribuciones muestrales de estadísticos (media, proporción y varianza) son la base de los mecanismos de la inferencia estadística:

1. La primera conclusión se puede resumir en el cuadro que aparece en seguida:

La media de la distribución muestral de la	Media	Es igual a la	Media		
	Proporción		Proporción	De la	Población
	Varianza modificada		Varianza modificada		

en símbolos:

$$\mu_X = E(\overline{X}) = \mu$$
$$\mu_p = E(\dot{p}) = \pi$$
$$\mu_{S_m^2} = E(S_m^2) = \pi_m^2$$

Esta última característica de la varianza modificada hace que sea un estimador insesgado, mientras que la varianza es un estimador sesgado porque no posee esta característica.

2. Existe una relación entre la desviación estándar de la población y la desviación estándar de la distribución muestral (el error estándar) y que es:

Error estándar de la	Símbolo con el que se le representa	Fórmula si se calcula a partir de la distribución muestral	Fórmula si se calcula a partir de la desviación estándar de la población	Fórmula simplificada (cuando no se utiliza el factor de corrección por población finita)
Media	$\sigma_{\bar{x}}$	$\pi_{\bar{x}} = \sqrt{\frac{\Sigma(\bar{X} - \mu)^2}{n}}$	$\pi_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N - n}{N - 1}}$	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$
Proporción	σ_p	$\sigma_p = \sqrt{\frac{\Sigma(P - \pi)^2}{n}}$	$\sigma_{\bar{x}} = \sqrt{\frac{\pi Q}{n}} \cdot \frac{N - n}{N - 1}$	$\sigma_p = \sqrt{\frac{\pi Q}{n}}$

3. **Teorema central del límite.** Si X es una variable aleatoria para la que se conocen su media μ y su varianza σ^2, la distribución muestral de la media tiende a ser normal con media $\mu_X = E(\bar{x}) = \mu$ y desviación estándar $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N - n}{N - 1}}$ conforme aumenta el tamaño de la muestra.

Si X es una variable aleatoria para la que se conocen su proporción π y su varianza σ^2, la distribución muestral de la proporción tiende a ser normal con proporción $\mu_p = E(p) = \pi$ y desviación estándar $\sigma_{\bar{x}} = \sqrt{\frac{\pi Q}{n}} \cdot \sqrt{\frac{N - n}{N - 1}}$ conforme aumenta el tamaño de la muestra.

Como se verá en capítulos posteriores, estas conclusiones son la base de los 2 principales mecanismos de la inferencia estadística: la estimación de parámetros y las pruebas de hipótesis.

7.6 Fórmulas del capítulo

7.2 Distribución muestral de la media

La media de la distribución muestral de las medias es igual a la media de la población o, dicho en otras palabras, el valor esperado de la media es igual a la media de la población:

$$E(X) = \mu_{\bar{X}} = \frac{\sum \bar{X}}{n} \quad (7.1)$$

La desviación estándar de la distribución muestral de la media (error estándar de la media):

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \quad (7.2)$$

7.2.3 Fórmula del error estándar de la media y factor de correción por población infinita

El error estándar de la media sin el factor de corrección por población finita:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \quad (7.3)$$

7.3 Distribución muestral de la proporción

La media de la distribución muestral de la proporción (valor esperado):

$$E(p) = \mu_p = \pi \quad (7.4)$$

La desviación estándar de la distribución muestral de la proporción (error estándar de la proporción):

$$\sigma_p = \sqrt{\frac{\pi Q}{n}} \sqrt{\frac{N-n}{N-1}} \quad (7.5)$$

7.3.3 Fórmula del error estándar de la proporción y factor de corrección por población finita

El error estándar de la proporción sin el factor de corrección por población finita:

$$\sigma_p = \frac{\sigma}{\sqrt{n}} = \frac{\sqrt{\pi Q}}{\sqrt{n}} = \sqrt{\frac{\pi Q}{n}} \quad (7.6)$$

7.4 Distribución muestral de la varianza

La varianza modificada de la población:

$$\sigma^2 = \frac{\sum (X-\mu)^2}{N-1} \quad (7.7)$$

La varianza modificada de la muestra:

$$s^2 = \frac{\sum (X-\mu)^2}{n-1} \quad (7.8)$$

7.7 Ejercicios adicionales

7.2 Distribución muestral de la media

1. Una empresa que fabrica tornillos produce un tipo de ellos que mide 5 cm de longitud con desviación estándar de 10 mm.

a) ¿Qué proporción de los tornillos que se fabrican miden entre 5 y 5.1 cm?
b) ¿Qué proporción de los tornillos fabricados mide entre 4.85 y 4.95 cm?
c) Si se eligen muchas muestras aleatorias de 25 bolsas, ¿se podría esperar que la media de esas muestras y la media poblacional fueran iguales? ¿Por qué?
d) ¿Cuál sería la forma de la distribución muestral de la media?
e) ¿Qué proporción de las medias muestrales estaría entre 5.0 y 5.1 cm?
f) ¿Qué proporción de las medias muestrales estaría entre 4.85 y 4.95 cm?

2. El promedio de tiempo de servicio de las cajas de una tienda de autoservicio es de 10 minutos con una desviación estándar de 2 minutos. Si se eligen muestras de 36 llamadas:

a) Calcule el error estándar de la media.
b) ¿Utilizó el factor de corrección por población finita? ¿Por qué?
c) ¿Qué proporción de los tiempos de servicio está entre 9 y 11 minutos?
d) ¿Qué proporción de las medias muestrales está entre 9 y 11 minutos?
e) Explique la razón de la diferencia entre las respuestas de los incisos *c*) y *d*).

Si ahora se toman muestras aleatorias de 100 elementos:

f) ¿Qué proporción de las medias muestrales estaría entre 9 y 11 minutos?
g) Explique la razón de la diferencia entre las respuestas de los incisos *d*) y *f*).

3. Una máquina envasadora llena botellas de refresco con 355 ml del líquido y el proceso de llenado tiene una desviación estándar de 25 ml; si se toman muestras de 36 botellas de refresco:

a) ¿Qué cantidad de líquido excedería 90% de las medias de las muestras?
b) ¿Hasta qué cantidad de líquido contendría 5% de las botellas llenadas con la menor cantidad de refresco?
c) ¿Dentro de qué cantidades de líquido caería 90% de las botellas alrededor de la media?

4. En la siguiente tabla se muestra el número de accidentes ocurridos dentro de la planta de una empresa textil durante el primer semestre del año.

Mes	Accidentes
Enero	3
Febrero	4
Marzo	2
Abril	1
Mayo	5
Junio	0

Encuentre:

a) La media de esta población.
b) La desviación estándar de esta población.
c) La distribución muestral de las medias para muestras de tamaño $n = 3$.
d) El valor esperado de la media.
e) El error estándar de la media, mediante:

- las medias de la distribución muestral de la media.
- la desviación estándar de la población.

f) ¿Se cumple que:

- $E(\overline{X}) = \mu$?
- $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}} \sqrt{\dfrac{N - n}{N - 1}}$?

g) Haga una gráfica de barras de la distribución muestral de la media agrupando los valores según su frecuencia de aparición, ¿qué forma aproximada tiene la gráfica?

5. Una tienda vende una marca de champú en sus 4 presentaciones, en la siguiente tabla se registró el número de botellas vendidas de cada una a lo largo de la semana.

Presentación	Ventas
A	10
B	4
C	8
D	6

Encuentre:

a) La media de esta población.
b) La desviación estándar de esta población.
c) La distribución muestral de las medias para muestras de tamaño $n = 2$.
d) El valor esperado de la media.
e) El error estándar de la media, mediante:

- las medias de la distribución muestral de la media.
- la desviación estándar de la población.

f) ¿Se cumple que:

- $E(\overline{X}) = \mu$?
- $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}} \sqrt{\dfrac{N - n}{N - 1}}$?

g) Haga una gráfica de barras de la distribución muestral de la media, agrupando los valores según su frecuencia de aparición, ¿qué forma aproximada tiene la gráfica?

6. A continuación se muestran los montos de 6 facturas que quedaron con adeudo en el departamento de cobranzas de una empresa en el mes de julio.

Cliente	Monto (miles)
Arteaga	3.5
Castro	7
Islas	17.5
Ríos	14
Suárez	10.5
Vázquez	21

Encuentre:

a) La media de esta población.
b) La desviación estándar de esta población.
c) La distribución muestral de las medias para muestras de tamaño $n = 4$.
d) El valor esperado de la media.
e) El error estándar de la media, mediante:

- las medias de la distribución muestral de la media.
- la desviación estándar de la población.

f) ¿Se cumple que:

- $E(\overline{X}) = \mu$?
- $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}} \sqrt{\dfrac{N - n}{N - 1}}$?

g) Haga una gráfica de barras de la distribución muestral de la media, agrupando los valores según su frecuencia de aparición, ¿qué forma aproximada tiene la gráfica?

7. En una oficina trabajan 5 telefonistas, a continuación se muestra el número de llamadas que cada una atiende durante un día.

Telefonista	Núm. de llamadas
Bertha	6
Diana	3
Fernanda	2
Lucía	4
Teresa	9

Encuentre:

a) La media de esta población.
b) La desviación estándar de esta población.
c) La distribución muestral de las medias para muestras de tamaño $n = 3$.
d) El valor esperado de la media.
e) El error estándar de la media, mediante:

- las medias de la distribución muestral de la media.
- la desviación estándar de la población.

f) ¿Se cumple que:

- $E(\overline{X}) = \mu$?
- $\sigma_{\bar{x}} = \dfrac{\sigma}{n} \sqrt{\dfrac{N - n}{N - 1}}$?

g) Haga una gráfica de barras de la distribución muestral de la media, agrupando los valores según su frecuencia de aparición, ¿qué forma aproximada tiene la gráfica?

8. El peso promedio de las latas de atún es de 42 g con una desviación de 15. Determine la probabilidad de que el peso promedio de una muestra de 80 latas sea:

a) Entre 38 y 44.
b) Menor de 40.
c) Mayor de 45.

9. El tiempo promedio en que se cobra a un cliente en las cajas de un supermercado es de 5.2 minutos, con una desviación estándar de 3.2. Determine la probabilidad de que en una muestra de 65 clientes el tiempo promedio para el cobro sea:

a) Entre 4 y 4.5 minutos.
b) Entre 5.5 y 6 minutos.

7.3 Distribución muestral de la proporción

10. De los productos que fabrica una máquina, 5% tienen defectos, si se seleccionan muestras de 50 productos:

a) Calcule el error estándar de la proporción.
b) ¿Qué proporción de las muestras tendría:
- entre 4 y 6% de productos defectuosos?
- más de 7% de productos defectuosos?
- menos de 4% o más de 5% de productos defectuosos?

11. En una encuesta sobre preferencias políticas se decide asumir que, si un candidato obtiene cuando menos 55% de los votos favorables, entonces será el candidato ganador en las elecciones. Si se toma una muestra aleatoria de 200 electores,¿cuál es la probabilidad de que se asuma como ganador de las elecciones a un candidato que:

a) tiene un porcentaje real de votos de 51%?
b) tiene un porcentaje real de votos de 55%?
c) tiene un porcentaje real de votos de 49%?
d) Determine las respuestas a los incisos *a*), *b*) y *c*) si se aumenta el tamaño de muestra a 500 y comente las diferencias.

12. Se tiene una población de 5 personas, de las cuales 2 tienen hijos; desde el punto de vista de los que tienen hijos encuentre:

a) La proporción de esta población.
b) La desviación estándar de esta población.
c) La distribución muestral de las proporciones para muestras de tamaño $n = 3$.
d) El valor esperado de la proporción.
e) El error estándar de la proporción, mediante:
- Las proporciones de la distribución muestral de la proporción.
- La desviación estándar de la población.

f) ¿Se cumple que:
- $E(p) = \mu_p = \pi$?
- $\sigma_p = \sqrt{\dfrac{\pi Q}{n}}\sqrt{\dfrac{N-n}{N-1}}$?

g) Haga una gráfica de barras de la distribución muestral de la proporción, agrupando los valores según su frecuencia de aparición, ¿qué forma aproximada tiene la gráfica?

13. En una reunión se encuentran 6 personas, de las cuales 3 fuman. Desde el punto de vista de los no fumadores encuentre:

a) La proporción de esta población.
b) La desviación estándar de esta población.
c) La distribución muestral de las proporciones para muestras de tamaño $n = 2$.
d) El valor esperado de la proporción.
e) El error estándar de la proporción, mediante:
- las proporciones de la distribución muestral de la proporción.
- la desviación estándar de la población.

f) ¿Se cumple que:
- $E(p) = \mu_p = \pi$?
- $\sigma_p = \sqrt{\dfrac{\pi Q}{n}}\sqrt{\dfrac{N-n}{N-1}}$?

g) Haga una gráfica de barras de la distribución muestral de la proporción, agrupando los valores según su frecuencia de aparición, ¿qué forma aproximada tiene la gráfica?

14. Se tiene una población de 7 alumnos, de los cuales 4 son alumnos regulares de la licenciatura de administración, desde el punto de vista de los no regulares encuentre:

a) La proporción de esta población.
b) La desviación estándar de esta población.
c) La distribución muestral de las proporciones para muestras de tamaño $n = 5$.
d) El valor esperado de la proporción.
e) El error estándar de la proporción, mediante:
- las proporciones de la distribución muestral de la proporción.
- la desviación estándar de la población.

f) ¿Se cumple que:
- $E(p) = \mu_p = \pi$?
- $\sigma_p = \sqrt{\dfrac{\pi Q}{n}}\sqrt{\dfrac{N-n}{N-1}}$?

g) Haga una gráfica de barras de la distribución muestral de la proporción, agrupando los valores según su frecuencia de aparición, ¿qué forma aproximada tiene la gráfica?

15. En una población de 8 artículos, de los cuales 5 pasaron el control de calidad, desde el punto de vista de los que pasaron el control encuentre:

a) La proporción de esta población.
b) La desviación estándar de esta población.
c) La distribución muestral de las proporciones para muestras de tamaño $n = 2$.

d) El valor esperado de la proporción.
e) El error estándar de la proporción, mediante:

- las proporciones de la distribución muestral de la proporción.
- la desviación estándar de la población.

f) ¿Se cumple que:

- $E(p) = \mu_p = \pi$?
- $\sigma_p = \sqrt{\frac{\pi Q}{n}} \sqrt{\frac{N - n}{N - 1}}$?

g) Haga una gráfica de barras de la distribución muestral de la proporción, agrupando los valores según su frecuencia de aparición, ¿qué forma aproximada tiene la gráfica?

16. Se tiene una población de 300 estudiantes de primaria y se sabe que 35% tiene caries; calcule la probabilidad de extraer de esta población una muestra de $n = 40$ elementos que tenga una proporción:

a) inferior a 0.15.
b) superior a 0.50.
c) entre 0.38 y 0.45.

17. En una población de 1 000 personas, 43% prefiere una marca específica de agua embotellada sobre la de la competencia. Calcule la probabilidad de extraer una muestra de $n = 35$ elementos tenga una proporción:

a) entre 0.30 y 0.38.
b) entre 0.52 y 0.63.

18. Del total de envíos de una empresa de paquetería, 27% presenta algún retraso en la entrega; calcule la probabilidad de que en una muestra de 60 elementos tenga una proporción:

a) inferior a 0.13.
b) superior a 0.34.
c) entre 0.30 y 0.40.

19. Del total de alumnos que egresan de la especialidad de mercadotecnia, 57% acreditan el dominio de algún idioma; calcule la probabilidad de que en una muestra de 43 elementos tenga una proporción:

a) entre 0.35 y 0.44.
b) entre 0.42 y 0.60.

7.4 Distribución muestral de la varianza

20. En la siguiente tabla se muestra el número de faltas que tuvieron las 6 telefonistas de un centro de atención al cliente durante el último mes.

Telefonista	Faltas
Bertha	0
Carla	4
Daniela	2
Isela	1
Laura	5
Nadia	4

Determine:

a) La varianza de la población.
b) La varianza modificada de la población.
c) La distribución muestral de la varianza para $n = 3$.
d) La distribución muestral de la varianza modificada para $n = 3$.
e) El valor esperado de la varianza.
f) El valor esperado de la varianza modificada.
g) ¿Se cumple que $(E(S^2) = \mu_{S^2} \neq \sigma^2)$?
h) ¿Se cumple que $(E(S_m^2) = \mu_{S_m^2} = \sigma_m^2)$?

21. En un conjunto de consultorios médicos trabajan 8 especialistas; a continuación se muestra el número de consultas que atendieron durante la última semana.

Especialista	Consultas
Cardiólogo	4
Dermatólogo	12
Fisioterapeuta	2
Ginecólogo	8
Neurólogo	2
Oftalmólogo	18
Radiólogo	22
Psicólogo	10

Determine:

a) La varianza de la población.
b) La varianza modificada de la población.
c) La distribución muestral de la varianza para $n = 2$.
d) La distribución muestral de la varianza modificada para $n = 2$.
e) El valor esperado de la varianza.
f) El valor esperado de la varianza modificada.
g) ¿Se cumple que $(E(S^2) = \mu_{S^2} \neq \sigma^2)$?
h) ¿Se cumple que $(E(S_m^2) = \mu_{S_m^2} = \sigma_m^2)$?

22. En un parque de diversiones se venden 4 tipos de entradas; a continuación se muestra cuántas se vendieron durante los últimos 15 días.

Tipo	Ventas
Adulto	90
Familiar	45
Infantil	65
Promocional	80

Determine:

a) La varianza de la población.
b) La varianza modificada de la población.

c) La distribución muestral de la varianza para $n = 3$.
d) La distribución muestral de la varianza modificada para $n = 3$.
e) El valor esperado de la varianza.
f) El valor esperado de la varianza modificada.
g) ¿Se cumple que $(E(S^2) = \mu_{S^2} \neq \sigma^2)$?
h) ¿Se cumple que $(E(S_m^2) = \mu_{S_m^2} = \sigma_m^2)$?

23. Un café internet renta las 6 computadoras con las que cuenta; en la siguiente tabla se muestra el número de minutos que las rentaron los clientes del pasado día.

Computadora	Minutos
A	60
B	180
C	210
D	30
E	120
F	150

Determine:

a) La varianza de la población.
b) La varianza modificada de la población.
c) La distribución muestral de la varianza para $n = 4$.
d) La distribución muestral de la varianza modificada para $n = 4$.
e) El valor esperado de la varianza.
f) El valor esperado de la varianza modificada.
g) ¿Se cumple que $(E(S^2) = \mu_{S^2} \neq \sigma^2)$?
h) ¿Se cumple que $(E(S_m^2) = \mu_{S_m^2} = \sigma_m^2)$?

24. En una granja pequeña se tienen 7 vacas, las cuales se ordeñan diariamente. En la siguiente tabla se muestra el número de litros que produjo cada una de ellas.

Vaca	L de leche
A	18
B	12
C	3
D	21
E	33
F	27
G	24

Determine:

a) La varianza de la población.
b) La varianza modificada de la población.
c) La distribución muestral de la varianza para $n = 5$.
d) La distribución muestral de la varianza modificada para $n = 5$.
e) El valor esperado de la varianza.
f) El valor esperado de la varianza modificada.
g) ¿Se cumple que $(E(S^2) = \mu_{S^2} \neq \sigma^2)$?
h) ¿Se cumple que $(E(S_m^2) = \mu_{S_m^2} = \sigma_m^2)$?

capítulo 8

Estimación de parámetros

Sumario

8.1 Estimaciones por punto y estimaciones por intervalo
8.2 Error de muestreo y errores que no son de muestreo
8.3 Propiedades de los estimadores
8.4 Estimación de una media con muestras grandes
8.4.1 Los 3 elementos de una estimación por intervalo
8.4.2 Estimador y parámetro estimado
8.4.3 Cuándo sí se puede utilizar la distribución normal para hacer estimaciones de parámetros, y cuándo no
8.4.4 Determinación del tamaño de muestra necesario para estimar una media
8.5 Comparación de la estimación de parámetros con muestras grandes y muestras pequeñas
8.5.1 Distribución *t* de Student, su tabla de áreas y Excel
8.6 Estimación de una media con muestras pequeñas
8.6.1 La población se distribuye de forma normal y se conoce la desviación estándar de la población: estadístico de prueba, *z*
8.6.2 La población se distribuye de forma normal pero no se conoce la desviación estándar de la población: estadístico de prueba, *t* de Student
8.6.3 La población no se distribuye de forma normal
8.7 Estimación de una proporción
8.7.1 Determinación del tamaño de muestra para estimar una proporción
8.8 Otros intervalos de confianza
8.8.1 Intervalos de confianza para la diferencia entre 2 medias poblacionales
8.8.2 Intervalos de confianza para la diferencia entre 2 proporciones poblacionales
8.8.3 Intervalos de confianza para el total de una población a partir de una media
8.8.4 Intervalos de confianza para el total de una población a partir de una proporción
8.9 Resumen
8.10 Excel: Uso de Excel para construir intervalos
8.11 Fórmulas del capítulo
8.12 Ejercicios adicionales

Estimar parámetros. Inferir el valor de la correspondiente medida de la población (parámetro) mediante una medida de una muestra (un estadístico).

Estimar parámetros es utilizar una medida de una muestra (un estadístico) para inferir el valor de la correspondiente medida de la población (parámetro); en este proceso, el estadístico se convierte en estimador.

En las 3 secciones primeras se revisarán diversos conceptos importantes para este tema: la diferencia entre **estimación por punto** y **estimación por intervalo**, el error de muestreo y las propiedades de los estimadores.

En las secciones del 8.4 al 8.8 se estudiarán los diversos aspectos y procedimientos estadísticos para la estimación de una media poblacional, incluida la determinación del tamaño de la muestra necesario para hacer la estimación y, finalmente, en las secciones del 8.10 al 8.12 se repasan las técnicas que se utilizan para hacer estimaciones por intervalo de proporciones y de varianzas.

8.1 Estimaciones por punto y estimaciones por intervalo

Estimación por punto. Utilizar un solo valor para estimar el parámetro.
Estimación por intervalo. Utilizar un rango de valores o intervalo.

Es posible hacer estimaciones de parámetros de 2 maneras: por punto y por intervalo. La **estimación por punto** consiste en utilizar un solo valor para estimar el parámetro, mientras que en la **estimación por intervalo** se utiliza un rango de valores o intervalo; así, una estimación por punto podría ser: "se estima que el promedio mensual de los ingresos de las familias de la zona metropolitana de Monterrey es de $4 500". Una estimación por intervalo podría ser: "se estima, con una confianza de 95% de estar en lo correcto, que el promedio de los ingresos de las familias de la zona metropolitana de la ciudad de Monterrey está entre $4 500 y $4 550".

El tipo de estimación que más se utiliza es la estimación por intervalo porque, como se aprecia en los ejemplos del párrafo anterior y como se verá con detalle más adelante, se le pueden asociar criterios de probabilidad ("...con una confianza de 95%...") que no es posible asociar con las estimaciones puntuales. Esto

es así porque las estimaciones de parámetros se hacen con base en las distribuciones muestrales que, como ya vimos, cuando n tiende a infinito poseen distribución normal y, siendo ésta una distribución continua, no es posible asociar probabilidades a valores individuales, en tanto que sí se puede calcular la probabilidad de ocurrencia de un determinado rango o intervalo de valores.

En el ejemplo anterior de una estimación por intervalo se pueden identificar 3 elementos:

1. El nivel de confianza: "...con una confianza de 95% de estar en lo correcto...". A este nivel de confianza se le denomina también el *coeficiente de confianza*.
2. El intervalo: "...entre $4 500 y $4 550." A los intervalos como éste se les suele llamar, precisamente, *intervalos de confianza* y se definen por el límite superior del intervalo y el límite inferior del intervalo.
3. La mención de que se trata de la estimación de un parámetro: "se estima...que el PROMEDIO de los ingresos de las familias de la zona metropolitana de la ciudad de Monterrey..."

Es importante tener presentes estos 3 elementos porque son indispensables para redactar la estimación, una vez que se realizaron las actividades necesarias.

En las secciones siguientes se revisan 2 temas importantes relacionados con la estimación de parámetros: el error de muestreo y las propiedades de los estimadores para, en el resto del capítulo, analizar ejemplos de los principales casos de estimaciones de parámetros:

a) de una media aritmética,
b) de la diferencia entre 2 medias,
c) de una proporción,
d) de la diferencia entre 2 proporciones de:

- una varianza.
- un total, a partir de una media y a partir de una proporción.

8.2 Error de muestreo y errores que no son de muestreo

Como las estimaciones de parámetros se hacen con muestras, la probabilidad de que el estadístico muestral sea exactamente igual al parámetro poblacional que se estima es 0, ya que las estimaciones suelen hacerse a través de distribuciones continuas, principalmente la normal. Por ello, al extraer una muestra, lo más probable es que el estadístico muestral no sea igual al correspondiente parámetro pero, al mismo tiempo, es altamente probable que sea un valor muy cercano ya que, como se desprende de la normal, la gran mayoría de los estadísticos muestrales se acercan al verdadero valor poblacional, sin embargo, es necesario tener en cuenta que al hacer inferencias estadísticas se incurre en un **error de muestreo** que puede definirse como la diferencia entre el estadístico muestral que se utiliza para la inferencia y el valor verdadero del parámetro correspondiente. Por otra parte, existen **errores que no se deben al muestreo** y que se deben a la mala aplicación de los procedimientos. Por ejemplo, puede haber errores al medir las variables, al transcribir datos de un documento a otro o al realizar operaciones con los datos obtenidos. Por supuesto la recomendación sobre este tipo de errores es que hay que esforzarse para evitarlos.

Error de muestreo. Diferencia entre el estadístico muestral que se utiliza para la inferencia y el valor verdadero del parámetro correspondiente.
Errores que no se deben al muestreo. Son aquellos que se deben a la mala aplicación de los procedimientos.

8.3 Propiedades de los estimadores

Ya se mencionó que la media aritmética y la proporción son medidas muy importantes en estadística, entre otras razones porque son muy útiles en la práctica, son medidas sencillas de calcular y, por lo general, todas las personas tienen una muy buena apreciación intuitiva de ellas; son además estadísticamente relevantes porque tienen propiedades que otras medidas no reúnen que las hacen especialmente útiles para hacer inferencias estadísticas. Estas propiedades son:

1. Ausencia de sesgo. En el capítulo anterior se definió que la media es un estimador insesgado porque la media de la distribución muestral es igual a la media de la población o, en otras palabras, que el valor esperado de la media es igual a la media de la población o, en símbolos: $\mu_{\overline{X}} = E(\overline{X}) = \mu$.

Como también se cumple que $\mu_p = E(p) = \pi$, entonces la proporción es también un estimador insesgado porque la media de la distribución muestral de la proporción es igual a la proporción de la población o, en otras palabras, porque el valor esperado de la proporción es igual a la proporción de la población.

Respecto a la varianza se vio que es un estimador sesgado porque $\mu_{S^2} = E(s^2) \neq \sigma^2$, es decir, la media de la distribución muestral de la varianza no es igual a la varianza de la población.

Pero se revisó también que la varianza modificada sí es un estimador insesgado de la varianza de la población, en donde la varianza modificada de una muestra es:

$$s_m^2 = \frac{\sum (X - \mu)^2}{n - 1}$$

Entonces se cumple que $\mu_{s_m^2} = E(s_m^2) \neq \sigma_m^2$.

Estimador consistente. Conforme aumenta el tamaño de la muestra, se incrementa la probabilidad de que el valor del estimador se aproxime al valor del parámetro.

2. Consistencia. Se dice que un **estimador** es **consistente** si, conforme aumenta el tamaño de la muestra, se incrementa la probabilidad de que el valor del estimador se aproxime al valor del parámetro. En otras palabras, si con el aumento de n aumenta la probabilidad de reducir el error de muestreo. La media aritmética y la proporción son estimadores consistentes.

Estimador eficiente. Es aquel con el menor error muestral, es decir, es aquel con la menor desviación estándar de su distribución muestral.

3. Eficiencia. La eficiencia se mide en términos comparativos. Se dice que, si se utilizan 2 estadísticos para estimar un parámetro, el **estimador** más **eficiente** es el que tiene el menor error muestral, es decir, es aquel con la menor desviación estándar de su distribución muestral. Se da una mayor eficiencia porque el menor error estándar hace que sea más probable que el valor del estimador sea más cercano al valor del parámetro. A manera de ejemplo, si se tiene una distribución simétrica y unimodal, la media y la mediana son iguales, y como la mediana es también un estimador insesgado de la media aritmética, se podría utilizar cualquiera de estos 2 estadísticos para estimar la media de la población; sin embargo, como el error estándar de la mediana es mayor que el de la media, entonces la media muestral es un estimador más eficiente que la mediana. La proporción es también un estimador eficiente.

Estimador suficiente. Es aquel que agota toda la información relevante que se puede extraer de una muestra.

4. Suficiencia. Un **estimador suficiente** es aquel que agota toda la información relevante que se puede extraer de una muestra. Al igual que antes, la media aritmética y la proporción son estimadores suficientes.

Resumiendo lo anterior, 2 de los más importantes estimadores que se utilizan en la inferencia estadística tienen las 4 propiedades que los hacen buenos estimadores: la media aritmética y la proporción. En la sección siguiente se revisa el procedimiento básico para estimar una media.

ejercicios 8.3 Introducción

1. Señale qué es *estimar parámetros*.
2. Explique la diferencia entre estimador por punto y estimador por intervalo.
3. ¿Qué son los errores de muestreo?
4. ¿Qué son los errores que no se deben al muestreo?
5. ¿Qué es un estimador insesgado?
6. ¿Qué es un estimador consistente?
7. ¿Qué es un estimador eficiente?
8. ¿Qué es un estimador suficiente?
9. ¿Cuáles estadísticos reúnen las 4 características de los buenos estimadores?

8.4 Estimación de una media con muestras grandes

Se explica en esta sección el procedimiento básico para estimar parámetros con muestras grandes ($n \geq 30$) y en secciones posteriores se hace una comparación entre este procedimiento y el que se debe utilizar con muestras pequeñas y se presentan diversos ejemplos.

El procedimiento básico para estimar una media con un intervalo de confianza consiste en obtener una muestra a partir de la cual se puede calcular la media de dicha muestra y, además, se puede utilizar como la mejor aproximación del verdadero valor de la media poblacional; el intervalo se construye alrededor de la media sumándole y restándole una cantidad que se forma por 2 elementos:

1. Un valor de z que representa el nivel de confianza y se basa en la distribución normal; por ejemplo, si se desea una confianza de 95%, sabemos, por la tabla de áreas bajo la curva normal, que a 1.95 desviaciones estándar (z) a la derecha y a la izquierda de la media se encuentra 95% de todos los valores posibles de la distribución muestral.

2. Como la medida de la dispersión de la distribución muestral de la media es el error estándar, la multiplicación de éste por el valor de z proporciona los límites dentro de los cuales se espera que esté el verdadero parámetro de la población.

Este procedimiento se puede resumir en forma simbólica de la siguiente manera:

$$\overline{X} \pm z\sigma_{\overline{X}} \tag{8.1}$$

Es decir, que la estimación:

- Se construye alrededor de la media de la muestra.
- Como sabemos que la distribución muestral de la media tiende a ser normal cuando el tamaño de la muestra hace lo propio a infinito; cuando se trabaja con muestras grandes se utiliza la distribución normal, con el valor de z, para incluir los cálculos de la probabilidad.
- Se utiliza el error estándar (la desviación estándar de la distribución muestral) de la media para medir la dispersión de las medias muestrales.

Se revisan en seguida 2 ejemplos.

EJEMPLO 8.1

Como prueba de un nuevo alimento para perros se revisan las ventas durante un mes en tiendas de autoservicio; los resultados de una muestra de 36 tiendas indican ventas promedio de \$12 000 por tienda con desviación estándar de \$800. Haga una estimación de intervalo con nivel de confianza de 95% para el promedio real de ventas para este nuevo alimento para perros.

Solución: Se tiene aquí que

$$\overline{X} = 12\,000$$
$$s = 800$$

Y para un nivel de confianza de 95%, si se busca en el cuerpo de la tabla de áreas bajo la curva normal el valor de z que aísle 47.5% del área a la derecha de la media, se encuentra que $P(-1.96 \leq z \leq 1.96) = 0.95$ y el intervalo con 95% de confianza:

$$\overline{X} \pm z\sigma_{\overline{X}} = \overline{X} \pm zs_{\overline{X}} = 12\,000 \pm 1.96\left(\frac{800}{\sqrt{36}}\right)$$

$$= 12\,000 \pm 1.96\frac{800}{6} = 12\,000 \pm 261.33$$

Así, el intervalo de confianza es \$11 738.67 a \$12 261.33 y se interpretaría afirmando que se tiene una confianza de 95% de estar en lo correcto al afirmar que, en la población total de tiendas de autoservicio, el promedio de ventas por tienda del nuevo alimento para perros está entre \$11 738.67 y \$12 261.33.

Nótese que en este ejemplo se utilizó la desviación estándar de la muestra, s, en vez de la correspondiente a la población, σ, para calcular el error estándar de la media, esto fue posible porque el tamaño de la muestra es relativamente grande ($n > 30$).

Obsérvese también que no se utilizó el factor de corrección por población finita para calcular el error estándar, se hizo así por 2 razones principales: la primera de ellas es que no se conoce N, el tamaño de la población, y la segunda es que se asume que esta N es lo suficientemente grande como para que haga que el valor del factor sea cercano a 1 y, por lo tanto, que su uso se haga innecesario.

Por otro lado, como los intervalos de confianza se construyen alrededor de la media de la muestra, en realidad se pueden construir una gran cantidad de intervalos diferentes, dependiendo del valor específico que tenga la media de la muestra particular que se obtiene, ya que, como se vio en el tema anterior sobre distribuciones muestrales, aunque lo más probable es que la media de una muestra aleatoria se acerque a la verdadera media de la población, la probabilidad de que sea exactamente igual es prácticamente de 0. Éste es el error muestral que se asume al hacer estimaciones.

EJEMPLO 8.2

De 50 000 peones de la construcción que laboran en el Distrito Federal, se tomó una muestra aleatoria de 400 y se investigó su ingreso diario. En la tabla siguiente se muestran los resultados. Construya un intervalo de confianza de 90% para el ingreso diario de la población total de peones.

Ingreso diario X	Número de peones f
50 a menos de 60	50
60 a menos de 70	80

(*continúa*)

(*continuación*)

Ingreso diario X	Número de peones f
70 a menos de 80	150
80 a menos de 90	100
90 a menos de 100	20
Total	400

Solución: En primer lugar, se calculan la media y la desviación estándar de la muestra:

X	F	Pm	$f \cdot Pm$	$(X - \bar{X})^2$	$f(X - \bar{X})^2$
50 a menos de 60	50	55	2 750	361	18 050
60 a menos de 70	80	65	5 200	81	6 480
70 a menos de 80	150	75	11 250	1	150
80 a menos de 90	100	85	8 500	121	12 100
90 a menos de 100	20	95	1 900	441	8 820
Totales	400		29 600		45 600

Así:

$$\bar{X} = \frac{29\,600}{400} = 74$$

$$s = \sqrt{\frac{45\,600}{400}} = \sqrt{114} = 10.68$$

De donde,

$$s_{\bar{X}} = \frac{10.68}{\sqrt{400}}\sqrt{\frac{N-n}{N-1}} = \frac{10.68}{20}\sqrt{\frac{50\,000 - 400}{50\,000 - 1}}$$

$$= 0.534(0.996) = 0.532$$

Y, como $P(-1.645 \leq z \leq 1.645) = 0.90$, el intervalo:

$$\bar{X} \pm z\sigma_{\bar{X}} = \bar{X} \pm zs_{\bar{X}} = 74 \pm 1.645(0.532) = 74 \pm 0.87$$

O sea, de \$73.13 a \$74.87; por ello, se puede afirmar, con una confianza de 90% de estar en lo correcto, que el ingreso diario de los peones de construcción del D.F. está entre \$73.13 y \$74.87.

Sobre estos procedimientos para estimar una media poblacional a partir de una media muestral, es necesario revisar varias consideraciones de las cuales se desprenden diferencias importantes en los procedimientos y los resultados de las estimaciones por intervalo; estas consideraciones se pueden dividir en 3 grandes categorías:

1. Consideraciones sobre cómo se construye la estimación por intervalo (los elementos de una estimación por intervalo).
2. Consideraciones sobre estimaciones de otros parámetros aparte de la media poblacional.
3. Consideraciones sobre cuándo sí y cuándo no es apropiado utilizar la distribución normal en la estimación (muestras grandes y muestras pequeñas).

En las secciones siguientes se aborda cada uno de estos temas y se incluyen ejemplos ilustrativos.

8.4.1 Los 3 elementos de una estimación por intervalo

Lo primero que es importante notar es que el intervalo se construye alrededor (±) de la media muestral, el *estadístico*. Además, el intervalo se abre (±) de acuerdo con el nivel de confianza especificado, z, y al error estándar del estimador, en este caso, el error estándar de la media, $s_{\bar{X}}$. En otras palabras, el intervalo de confianza tiene 3 elementos:

1. El estadístico, en los ejemplos anteriores la media muestral, alrededor del cual se construye el intervalo.
2. El nivel de confianza especificado, que se utiliza para determinar el valor de z, la desviación estándar de la distribución normal estandarizada, que es la que permite medir las probabilidades de acuerdo con esta distribución.
3. El error estándar de la media, $s_{\bar{x}}$, que aquí se calculó a partir de la desviación estándar de la muestra.

Con estos mismos 3 elementos se construyen diferentes intervalos de confianza dependiendo de diversas circunstancias que son el tema del resto de este capítulo.

Antes de pasar al siguiente apartado, se hace un resumen del procedimiento a seguir para estimar una media que puede extenderse para estimar cualquier otro parámetro o combinación de ellos:

1. Determinar el estadístico muestral (la media, proporción, etc., de la muestra).
2. Determinar el valor de z (o de otras distribuciones como la t de Student, que se analizará más adelante) correspondiente al nivel de confianza que se desea.

3. Calcular el error estándar del estadístico.
4. Determinar el intervalo.
5. Interpretar el intervalo.

Para estudiantes de ciencias sociales (y, en realidad, de todas las áreas del conocimiento), este último paso es de especial importancia porque implica entender y explicar en términos de la realidad lo que técnicamente es un intervalo de confianza.

8.4.2 Estimador y parámetro estimado

En los ejemplos anteriores se estimó la media de una población a partir de la media de una muestra. Se pueden construir intervalos de confianza similares para diferentes parámetros o combinaciones de ellos, con los mismos 3 elementos. Entre los principales, es posible hacer estimaciones por intervalo para:

- una media aritmética,
- una proporción,
- un total a partir de una media,
- un total a partir de una proporción,
- la diferencia entre 2 medias,
- la diferencia entre 2 proporciones,
- una varianza, entre otros.

Tal como se menciona antes, en este capítulo se revisan algunos ejemplos.

8.4.3 Cuándo sí se puede utilizar la distribución normal para hacer estimaciones de parámetros, y cuándo no

Se mencionó ya que se utiliza la z de la distribución normal para hacer estimaciones de parámetros cuando la muestra es relativamente grande, sin importar ninguna otra consideración, y se fijó que el tamaño de muestra a partir del cual se considera que una muestra es grande es 30. Sin embargo, cuando la muestra es pequeña, es decir, cuando $n \leq 30$ y, además:

1. La distribución de la variable en la población es normal.
2. No se conoce la desviación estándar de la población y se utiliza la de la muestra, s.

Entonces se debe utilizar la distribución t de Student y no la z de la distribución normal. Se enfatiza esto porque es importante tenerlo en cuenta; en la sección 8.6 se abunda sobre ello y se dan más detalles y ejemplos.

8.4.4 Determinación del tamaño de muestra necesario para estimar una media

Se puede determinar el tamaño de muestra necesario para estimar una media partiendo del mecanismo de estimación y se hace incluyendo el factor de corrección por población finita, o no.

8.4.4.1 Cuando no se incluye el factor de corrección por población finita

En este caso, se parte de la fórmula que resume el procedimiento de estimación:

$$\overline{X} \pm z\sigma_{\overline{X}}.$$

Si consideramos que $z\sigma^2$ es un margen de error alrededor de la media de la muestra, entonces podemos escribir $e = z\sigma_{\overline{X}}$ y si sustituimos $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$ en esta última expresión, se obtiene:

$$e = z\frac{\sigma}{\sqrt{n}}$$

Despejando n, el tamaño de la muestra de la expresión anterior se llega a:

$$n = \left(\frac{z\sigma}{e}\right)^2 \qquad \textbf{(8.2)}$$

EJEMPLO 8.3

El gerente de personal de una empresa grande desea estimar, con una confianza de 95%, el tiempo promedio de capacitación que recibieron los empleados de la compañía y desea que el error de la estimación no rebase los 30 minutos (0.5 horas). Con base en resultados de estudios anteriores, estima que la desviación estándar del tiempo de capacitación de los empleados es de 3 horas. El tamaño mínimo de muestra para este estudio es:

$$n = \left(\frac{z\sigma}{e}\right)^2 = \left(\frac{1.96(3)}{0.5}\right)^2 = 138.30$$

Por lo que el tamaño mínimo de la muestra debe ser $n = 139$. Nótese aquí que el redondeo se hace hacia arriba, sin importar cuál sea la fracción decimal.

8.4.4.2 Cuando sí se incluye el factor de corrección por población finita

En este caso, la expresión que resume el procedimiento para realizar estimaciones de una media es:

$$e = z\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}.$$

Despejando n:

$$\frac{e\sqrt{n}}{z\sigma} = \sqrt{\frac{N-n}{N-1}}$$

$$\frac{e^2 n}{z^2\sigma^2} = \frac{N-n}{N-1}$$

$$e^2 n(N-1) = (N-n)z^2\sigma^2$$

$$e^2 nN - e^2 n = Nz^2\sigma^2 - nz^2\sigma^2$$

$$e^2 nN - e^2 n + nz^2\sigma^2 = Nz^2\sigma^2$$

$$n(e^2 N - e^2 + z^2\sigma^2) = Nz^2\sigma^2$$

$$n = \frac{Nz^2\sigma^2}{e^2 N - e^2 + z^2\sigma^2} = \frac{Nz^2\sigma^2}{e^2(N-1) + z^2\sigma^2}$$

$$n = \frac{Nz^2\sigma^2}{e^2(N-1) + z^2\sigma^2} \tag{8.3}$$

EJEMPLO 8.4

Un gerente de personal de una empresa que tiene 3 000 empleados desea estimar el tiempo promedio que les lleva a sus trabajadores trasladarse entre sus casas y la empresa. Desea construir un intervalo de confianza de 95% y que la estimación esté dentro de un minuto alrededor de la media verdadera; a través de una muestra piloto se sabe que la varianza de los tiempos es de 30 minutos. Determine el tamaño de la muestra que se requiere para hacer la estimación.

Solución: Aquí,

$N = 3\,000$
$z = 1.96$
$e = 1$
$s^2 = 30.$

De donde

$$n = \frac{3\,000(1.96)^2(30)}{1^2(3\,000-1) + 1.96^2(30)}$$

$$= \frac{345\,744}{2\,999 + 115.248} = \frac{345\,744}{3\,114.248} = 111.02$$

Que se redondea a 112.

ejercicios 8.4 Estimación de una media con muestras grandes

1. ¿Cuáles son los 3 elementos que se utilizan para estimar parámetros?

2. Describa el procedimiento para estimar parámetros.

3. ¿Cuáles son los parámetros sobre los que más comúnmente se realizan estimaciones?

4. Explique de dónde se desprenden las fórmulas para calcular tamaños de muestra para realizar estimaciones de parámetros.

5. Se tomó una muestra de 35 empleados de una empresa que, en promedio, tienen un salario diario de $133, con una desviación estándar muestral de $6. Haga una estimación de intervalo con nivel de confianza de 95% para el promedio de salario diario del total de trabajadores de la empresa.

6. Una cadena de comida rápida registró el número de clientes que recibió durante un día en una muestra aleatoria de 43 de los restaurantes que tiene en el país; se encontró que en promedio se recibieron 107 clientes diarios con una desviación estándar de 23. Haga una estimación de intervalo con nivel de confianza de 90% para el promedio de clientes que se reciben en el total de restaurantes.

7. Se tomó una muestra aleatoria de 400 llantas producidas en una planta, y se encontró que la vida útil promedio era de 43 000 kilómetros con una desviación estándar de 1 300. Haga una estimación por intervalo para la vida útil promedio del total de llantas producidas con niveles de confianza:

a) de 85%.
b) de 95%.

8. En 33 oficinas postales elegidas al azar se recibieron en promedio 1 357 cartas durante un día específico, con una desviación estándar de 226. Haga una estimación de intervalo con un nivel de confianza de 92% para el promedio de las cartas recibidas en ese mismo día en todas las oficinas.

9. De los 15 000 alumnos que asisten a una institución universitaria, se preguntó a 120 cuánto tiempo les tomaba llegar a la escuela. En la siguiente tabla se muestran los resultados.

Tiempo (minutos) *X*	Número de alumnos *f*
0 a menos de 20	18
20 a menos de 40	30
40 a menos de 60	41
60 a menos de 80	19
80 a menos de 100	12
Total	120

Haga una estimación de intervalo con nivel de confianza de 90% para el tiempo promedio de traslado del total de estudiantes.

10. Una cadena de pizzerías con servicio a domicilio cuenta con 120 tiendas a lo largo del país; se tomó una muestra de 50 tiendas y se registró el número de pedidos que recibieron durante un día, con los resultados que se muestran a continuación. Haga una estimación por intervalo con nivel de confianza de 80% para promedio de ventas en el total de tiendas en ese día.

Pedidos *X*	Número de tiendas *f*
20 a menos de 40	4
40 a menos de 60	7
60 a menos de 80	22
80 a menos de 100	10
100 a menos de 120	6
120 a menos de 140	1
Total	50

11. De 2 000 alumnos que asisten a una escuela secundaria, se tomó una muestra de 200 a los que se les preguntó el número de horas que dedican a la semana para practicar algún deporte. Haga una estimación de intervalo con un nivel de confianza de 85% para el número promedio de horas que dedica al deporte la población completa de alumnos.

Número de horas *X*	Número de alumnos *f*
2 a menos de 5	23
5 a menos de 8	43
8 a menos de 11	64
11 a menos de 14	50
14 a menos de 17	20
Total.	100

12. De los 300 días en un año en que funciona una imprenta, se tomó el número de impresiones en blanco y negro realizados durante 90 días, los resultados se muestran en la siguiente tabla. Haga una estimación de intervalo con un nivel de confianza de 95% para el promedio de impresiones en blanco y negro de todos los días.

Número de impresiones *X*	Número de días *f*
2 a menos de 10	3
10 a menos de 18	6
18 a menos de 26	20
26 a menos de 34	26
34 a menos de 42	16
42 a menos de 50	10
50 a menos de 58	7
58 a menos de 66	2
Total	90

Tamaño de la muestra sin incluir el factor de corrección por población finita

13. Los registros históricos de una cadena de tiendas muestran que la desviación estándar de las ventas por tienda es de $2 000. Se desea hacer una nueva estimación del promedio de ventas por tienda con un nivel de confianza de 99% y considere que el margen de error alrededor de la media no sea mayor que $500. Determine el tamaño necesario de la muestra.

14. Un inversionista considera la posibilidad de adquirir una tienda de regalos en una terminal de autobuses foráneos y desea estimar el promedio de ventas por cliente, con base en otras tiendas similares que ya posee, estima que la desviación estándar de las ventas por cliente es de $15. ¿Cuál es el tamaño mínimo de la muestra aleatoria que debe tomar para hacer la estimación si desea un nivel de confianza de 90% y un margen de error no mayor que $1?

15. Un restaurante desea conocer, con un nivel de confianza de 90%, la cantidad de clientes promedio que acuden entre 11 a.m. y 6 p.m. los lunes y considera que el error de la estimación no debe rebasar los 2 clientes; basándose en las muestras del resto de los días de la semana, estima que la desviación estándar es de 5 clientes. Calcule el tamaño de la muestra necesario para llevar a cabo dicha estimación.

16. Una embotelladora quiere saber el promedio de botellas que se reciclan bimestralmente con un nivel de confianza de 99%; si se considera que la desviación estándar del número de botellas recicladas cada bimestre es de 35 y se especifica que el error de la estimación no debe superar 7 botellas, calcule el tamaño de la muestra necesario para llevar a cabo dicha estimación.

17. Se sabe que la desviación estándar del nivel de ventas de cierto producto por tienda es de $350 y se supone que el nivel de ventas por establecimiento sigue una distribución normal. Determine el tamaño mínimo de la muestra requerida para estimar las ventas promedio por tienda con un margen de error inferior a $150 con una confianza de 95 por ciento.

Tamaño de la muestra incluyendo el factor de corrección por población finita

18. Un gerente de personal desea estimar el número promedio de horas de capacitación que toman los vendedores de la empresa, con un margen de error de 2 horas y un nivel de confianza de 95%; con base en estudios anteriores se sabe que la desviación estándar de las horas de capacitación de los vendedores es de 10 horas. Calcule el tamaño de muestra mínimo que se requiere, sabiendo que el número de vendedores de la empresa es de 2 000.

19. El departamento de salud desea estimar el promedio de veces que los alumnos de primaria de un distrito escolar acuden al médico anualmente, con un nivel de confianza de 93% y un margen de error de 2, se sabe que la desviación estándar es de 12. Calcule el tamaño de muestra mínimo que se requiere, sabiendo que el número de alumnos en el nivel primario del distrito son 18 320.

20. Una empresa desea estimar el promedio de trabajadores que requieren utilizar lentes, con un nivel de confianza de 90% y un margen de error de 2, se sabe que la desviación estándar es de 8. Calcule el tamaño de muestra mínimo que se requiere, sabiendo que el número de trabajadores de la empresa es de 3 225.

21. El departamento de finanzas quiere estimar la cantidad de litros de gasolina que utilizan los autos destinados al personal de ventas con un nivel de confianza de 96% y un margen de error de 2, se conoce que la desviación estándar es de 15. Calcule el tamaño de muestra mínimo que se requiere, sabiendo que el número de autos es de 350.

22. Un productor quiere estimar, con un nivel de confianza de 99% y un margen de error de 1, el promedio de manzanas por caja que llegan en mal estado al almacén de donde se surten todos los pedidos; se sabe que la desviación estándar es de 2.5. Calcule el tamaño de muestra mínimo que se requiere, sabiendo que el número de cajas de manzanas en cada trayecto es de 5 000.

8.5 Comparación de la estimación de parámetros con muestras grandes y muestras pequeñas

Dado el planteamiento del teorema central del límite que establece, como se dijo repetidamente, que la distribución muestral de los estadísticos tiende a ser normal cuando el tamaño de la muestra tiende a infinito, es fácil visualizar por qué con frecuencia se utiliza z, la desviación estándar de la distribución normal estándar (con media de 0 y desviación estándar de 1) para hacer estas inferencias estadísticas. De hecho, cuando se tiene una muestra grande, carece de importancia la forma de la distribución de la variable en la población: puede ser normal o no y aun así, su distribución muestral tiene forma normal. Además, tampoco importa que se conozca o no el verdadero valor de la desviación estándar de la población pues, si no se conoce, se puede utilizar la desviación estándar de la muestra para sustituir a la de la población en el cálculo del error estándar.

Resumiendo, se utiliza la distribución normal para hacer estimaciones de la media cuando se tienen muestras grandes, sin importar qué distribución tenga la variable en la población y sin importar si se co-

noce o no la desviación estándar de la población. Por lo general, se acepta que una muestra es "grande" si tiene cuando menos 30 elementos; en símbolos, si $n \geq 30$.

Por otra parte, cuando la muestra es pequeña (menor de 30) sólo se puede utilizar z si se sabe con certeza que la distribución de la variable en la población es normal y si, además, se conoce la desviación estándar de la población.

Existen otros 2 casos posibles para cuando se tiene una muestra pequeña; uno de ellos es cuando la distribución de la variable en la población es normal pero no se conoce la desviación estándar de la población; en este caso se utiliza una distribución conocida como *distribución* t *de Student*, en vez de la normal. En el apartado 8.5.1 de esta sección se presenta esta distribución y en la sección siguiente, la 8.6, se resuelven algunos ejemplos con esta distribución t de Student. El otro caso posible para muestras pequeñas es cuando la distribución de la variable en la población no es normal; aquí, independientemente de si se conoce o no la desviación estándar de la población, no se puede utilizar ni z ni t, aunque es posible que se pueden aplicar algunos de los métodos no paramétricos que se estudian en el capítulo 18. Se resumen estas consideraciones en el esquema que se presenta en el cuadro 8.1.

Cuadro 8.1 Inferencias con muestras grandes o pequeñas (cuándo usar z y cuándo t de Student al hacer estimaciones)

Tamaño de la muestra	Condiciones	Distribución a utilizar
$n \geq 30$	Sólo el tamaño de la muestra.	Normal, con z.
$n < 30$	1. La distribución de la variable en la población es normal. 2. Se conoce la desviación estándar de la población, σ.	Normal, con z.
$n < 30$	1. La distribución de la variable en la población es normal. 2. No se conoce la desviación estándar de la población y se utiliza la de la muestra, s.	t de Student.
$n < 30$	1. La distribución de la variable en la población no es normal.	Posiblemente procedimientos no paramétricos (capítulo 18).

8.5.1 Distribución *t* de Student, su tabla de áreas y Excel

Cuando se tienen muestras pequeñas, con menos de 30 elementos y no se conoce la desviación estándar de la población, pero la población se distribuye de manera normal, se puede utilizar como estadístico de prueba, la t de la distribución t de Student.

En este caso, se utiliza la desviación estándar de la muestra, s, para estimar la de la población y, por ello, el cálculo del error estándar de la media se convierte en:

$$s_{\bar{X}} = \frac{s}{\sqrt{n}}$$

Al sustituir esta estimación del error estándar en la fórmula que antes se utilizó para encontrar el valor de z, se llega a:

$$t = \frac{\bar{X} - \mu}{s_{\bar{X}}} = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$$

Haciendo esto se llega a una variable que no se distribuye como z, sino como otra distribución, la distribución t de Student, cuyas principales propiedades son las siguientes:

- Tiene media de 0.
- Es simétrica respecto a la media.
- No se trata de una sola distribución sino, más bien, de una familia de distribuciones, cada una de ellas definida por los denominados *grados de libertad* dados por $n - 1$. En símbolos $gl = n - 1$.
- En términos generales, esta distribución t es más achatada que la normal en el centro y sus colas son más altas que las de la normal pero, la distribución t, se aproxima a la normal conforme n aumenta.

Al igual que las probabilidades (áreas) que se obtienen a partir de la distribución normal, las probabilidades (áreas) asociadas con la distribución t de Student se suelen resumir en tablas, como la tabla 3 de la sección

"Apéndice de tablas" que se encuentra al final de este libro; dicha tabla se calculó a partir de la función "Distr.T.Inv" de Excel. En la sección siguiente se ilustra el uso de esta función y de la tabla.

8.6 Estimación de una media con muestras pequeñas

El procedimiento para estimar una media poblacional cuando se utilizan muestras pequeñas es prácticamente igual al que ya se vio para muestras grandes: se construye un intervalo alrededor de la media muestral sumándole y restándole la t (según el nivel de confianza que se desea) multiplicada por el error estándar. Como se ve en el cuadro 8.1, y como se explica más adelante, en caso de que:

- la distribución de la variable en la población es normal.
- no se conoce la desviación estándar de la población y se utiliza la de la muestra, s.

Entonces se debe utilizar la distribución t de Student y no la z de la distribución normal. Se revisan en seguida los diversos casos que se deben considerar al hacer estimaciones de parámetros con muestras pequeñas.

8.6.1 La población se distribuye de forma normal y se conoce la desviación estándar de la población: estadístico de prueba, *z*

Cuando la muestra es pequeña pero se cumplen estas 2 condiciones:

1. se sabe o se puede suponer que la variable se distribuye en forma normal en la población y
2. se conoce la desviación estándar de la población.

Entonces puede utilizar la distribución normal (z), tal como se ilustró antes.

EJEMPLO 8.5

La vida útil de los focos para iluminación de escenarios que produce una empresa tiene una desviación estándar de 40 horas, si se toma una muestra aleatoria de 25 focos y se encuentra que su vida útil promedio es de 835 horas, construya un intervalo de confianza de 95% para el promedio de vida útil de esos focos.

Solución:

$\overline{X} = 835$

$\sigma = 40$

De donde

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{25}} = \frac{40}{5} = 8$$

Y, como $P(-1.96 \leq z \leq 1.96) = 0.95$, el intervalo:

$$\overline{X} \pm z\sigma_{\overline{X}} = 835 \pm 1.96(8) = 835 \pm 15.68$$

Es decir, de 819.32 a 850.68, y, entonces, se estima con una confianza de 95% que el promedio de la vida útil de esos focos para iluminación de escenarios está entre 819.32 y 850.68 horas.

8.6.2 La población se distribuye de forma normal pero no se conoce la desviación estándar de la población: estadístico de prueba, *t* de Student

Ya se anotaron antes, en el diagrama del cuadro 8.1, las condiciones en las que se debe utilizar z y cuándo está t de Student en las estimaciones. Se presenta en seguida un ejemplo.

EJEMPLO 8.6

En una muestra de 10 comprobantes de compra de un supermercado, la compra promedio fue de \$114, con una desviación estándar de \$33; si las compras siguen una distribución aproximadamente normal, haga una estimación de intervalo con un nivel de confianza de 99% del promedio de compra del total de compras.

Solución: En este caso, como la muestra es menor de 30 pero se sabe que la distribución de la variable (monto de las compras) se distribuye de manera normal en la población, se puede utilizar la distribución t de Student para hacer la estimación.

El estadístico muestral, la media, es $\overline{X} = 114$.
El error estándar de la media:

$$s_{\overline{X}} = \frac{s}{\sqrt{n}} = \frac{33}{\sqrt{10}} = \frac{33}{3.16} = 10.44$$

El estadístico de prueba es, entonces, la t de Student y se tienen 9 grados de libertad:

$$gl = n - 1 = 10 - 1 = 9$$

Consultando la tabla de esta distribución t, la tabla de la parte interior de la contraportada, se busca la probabilidad en el extremo (el encabezado de las columnas) correspondiente a 0.005, resultado de 1 – 0.99 = 0.01 dividido entre 2, para ambos extremos, y en la intersección con el renglón correspondiente a los 9 grados de libertad; se encuentra aquí el valor $t = 3.25$ o, en símbolos,

$$P(-3.25 \leq t_9 \leq 3.25) = 0.99$$

Este mismo resultado utilizando la función Distr.T.Inv de Excel: =DISTR.T.INV(0.01,9), produce como resultado 3.249836, que es el mismo valor que se obtuvo a partir de la tabla, pero con mayor precisión.

Y, el intervalo:

$$\overline{X} \pm t_{gl} s_{\overline{x}} = 114 \pm 3.250(10.44) = 114 \pm 33.93$$
$$80.07 \leq \mu \leq 147.93.$$

El intervalo de confianza es \$80.07 a \$147.93 y se le interpretaría con la afirmación de que se tiene una confianza de 99% de estar en lo correcto al aseverar que, en la población total de compras, la compra promedio está entre \$80.07 y \$147.93.

EJEMPLO 8.7

Se realizó un estudio en 20 hogares elegidos aleatoriamente en una ciudad grande para medir la cantidad de agua que utilizaba por hora. Los resultados se muestran en la tabla siguiente.

Consumo en litros X	Hogares f
50 a menos de 70	1
70 a menos de 90	1
90 a menos de 110	2
110 a menos de 130	3
130 a menos de 150	7
150 a menos de 170	6
Total	20

Haga una estimación de intervalo con un nivel de confianza de 95% para la población total de hogares.

Solución: En primer lugar se calculan la media y la desviación estándar del consumo de agua por hora en la muestra. En la tabla siguiente se muestran los cálculos correspondientes.

Consumo en litros X	Hogares f	Pm	$f \cdot Pm$	$(X - \overline{X})^2$	$f(X - \overline{X})^2$
50 a menos de 70	1	60	60	5 184	5 184
70 a menos de 90	1	80	80	2 704	2 704
90 a menos de 110	2	100	200	1 024	2 048
110 a menos de 130	3	120	360	144	432
130 a menos de 150	7	140	980	64	448
150 a menos de 170	6	160	960	784	4 704
Total	20		2 640		15 520

$$\overline{X} = \frac{2\,640}{20} = 132$$

$$s = \sqrt{\frac{15\,520}{19}} = \sqrt{816.84} = 28.58$$

El error estándar de la media:

$$s_{\overline{X}} = \frac{s}{\sqrt{n}} = \frac{28.58}{\sqrt{20}} = \frac{28.58}{4.47} = 6.39$$

Ahora, como se sabe que la distribución del consumo de agua en la población sigue una distribución normal, se puede utilizar la distribución t de Student, con 19 grados de libertad: $gl = n - 1 = 20 - 1 = 19$.

En la tabla de esta distribución, $P(-2.093 \leq t_{19} \leq 2.093) = 0.95$.

De donde, el intervalo es:

$$\overline{X} \pm t_{gl} s_{\overline{x}} = 132 \pm 2.093(6.23) = 132 \pm 13.04$$

O sea, de 118.96 a 145.04 litros por hora y se interpretaría afirmando que se tiene una confianza de 95% de estar en lo correcto al aseverar que, en la población total de hogares en esa ciudad, el consumo promedio de agua por hora está entre 118.96 y 145.04 litros por hora.

8.6.3 La población no se distribuye de forma normal

En este caso, simplemente no es posible utilizar los mecanismos descritos hasta aquí, por lo que sólo quedan 2 alternativas:

1. aumentar el tamaño de la muestra para que sea de cuando menos 30 elementos y
2. considerar la posibilidad de utilizar alguno de los métodos no paramétricos como los que se analizan en el capítulo 18.

ejercicios 8.6 Estimación de una media con muestras pequeñas

Se conoce la desviación estándar de la población

1. El consumo de carne de res mensual en determinada comunidad tiene una desviación estándar de 1.2 kilogramos, se tomó una muestra aleatoria de 20 personas y se encuentra que el consumo promedio es de 5.3 kilogramos; construya un intervalo de confianza de 96% para el promedio de consumo de carne.
2. Se desea estimar el número de productos que adquieren los clientes de un supermercado; por estudios anteriores, se sabe que la desviación estándar del número de productos comprados es 9. Se tomó una muestra aleatoria de 15 amas de casa y se encontró que en promedio adquirieron 43 productos. Construya un intervalo de confianza de 98% para el promedio de compras por persona.
3. Una tienda departamental ofrece a sus clientes una tarjeta para realizar sus compras a crédito y se sabe que la desviación estándar de los saldos que no están al corriente de sus pagos es de $720. Se tomó una muestra aleatoria de 27 personas y se encontró que en promedio deben $2 217. Construya un intervalo de confianza de 90% para el promedio de los saldos en mora de los usuarios de la tarjeta.
4. En una pizzería, el tiempo que se tardan en cubrir una orden tiene una desviación estándar de 7.3 minutos, se tomó el tiempo que tardaron en cubrirse 19 órdenes elegidas al azar y se encontró que en promedio se tardaron 23.4 minutos. Construya un intervalo de confianza de 95% para el tiempo promedio para cubrir cada orden.
5. La cantidad de camarón diario que pesca un barco que opera en una cooperativa de pescadores de la costa de Sinaloa durante una temporada tiene una desviación estándar de 37.8 kilos; se registró la cantidad pescada en 25 días tomados aleatoriamente y se encontró que en promedio se pescan 567 kilos de camarón al día. Construya un intervalo de confianza de 93% para la cantidad de camarón capturado.

No se conoce la desviación estándar de la población pero se sabe que la variable se distribuye de manera normal

6. La duración promedio de una muestra de 12 focos fue de 4 005 horas, con desviación estándar de 210 horas; si se sabe que la duración de los focos producidos en general sigue una distribución normal, haga una estimación de intervalo con un nivel de confianza de 95% de la duración del total de focos.
7. Se repartieron por correo cupones de descuento alrededor de la ciudad, en 27 tiendas en promedio se hicieron válidos 32 cupones por día en cada una con una desviación estándar de 12; si se supone que el cambio de los cupones sigue una distribución aproximadamente normal, haga una estimación de intervalo con un nivel de confianza de 80% del número de cupones que se hacen válidos en el total de tiendas.
8. Se tomó una muestra de 20 velas que estuvieron encendidas hasta que se consumieron, se observó que en promedio se consumieron en 58 horas con una desviación estándar de 8. Si se supone que el ciclo de vida del total de velas tiene una distribución aproximadamente normal, haga una estimación de intervalo con un nivel de confianza de 90% de la duración total de la población de velas.
9. Se realizó una encuesta a 25 personas a las que se les mostró un nuevo modelo de reloj y se les preguntó cuál es el precio máximo que estarían dispuestas a pagar por él; en promedio el precio aceptado fue de $527 con una des viación estándar de $67. Si las respuestas siguen una distribución aproximadamente normal, haga una estimación por intervalo con un nivel de confianza de 95% del precio máximo promedio que los posibles clientes en general pagarían por el reloj.
10. Se preguntó a 15 estudiantes de segundo grado de secundaria el número de libros que leyó en el último año, en promedio leyeron 5 libros con una desviación estándar de 1.5. Haga una estimación de intervalo con los siguientes niveles de confianza para el promedio de libros que leyeron todos los estudiantes de segundo grado de secundaria durante ese último año:

    a) 90%.
    b) 98%.

8.7 Estimación de una proporción

Tal como se vio para la media aritmética, el intervalo de estimación se construye sumando al estadístico el producto de la z correspondiente al nivel de confianza por el valor del error estándar, o $\overline{X} \pm z\sigma_{\overline{X}}$. El procedimiento para estimar una proporción es igual, salvo que, por supuesto, se utilizan la proporción de la muestra y su correspondiente error muestral. En símbolos:

$$p \pm z\sigma_p \tag{8.4}$$

EJEMPLO 8.8

Se desea estimar, con una confianza de 99%, la proporción de alumnos de una universidad que acuden a sus instalaciones en su propio automóvil; se toma una muestra de 200 alumnos y se encuentra que 25 de ellos manifiestan tener automóvil.

Solución: La proporción de la muestra es $p = \frac{25}{200} = 0.125$

La z correspondiente a 99% de confianza es de 2.575, ya que: $P(-2.575 \geq z \leq 2.575) = 0.99$.

El error estándar es: $s_p = \sqrt{\frac{pq}{n}} = \sqrt{\frac{0.125(0.875)}{200}} = 0.0234$.

Y el intervalo:

$$p \pm zs_p = 0.125 \pm 2.575(0.0234) = 0.125 \pm 0.06$$

O sea, de 0.065 a 0.185.

Así, se estima con una confianza de 99% que la proporción de estudiantes que posee automóvil está entre 6.5 y 18.5 por ciento.

EJEMPLO 8.9

La gerencia de recursos humanos de una empresa quiere estimar, con un nivel de confianza de 96%, la proporción de los trabajadores de la empresa que cuentan con casa propia; de una muestra de 350 empleados, 50% dijo tener casa propia.

Solución: $p = 0.5$

$P(-2.054 \geq z \leq 2.054) = 0.96$

$$s_p = \sqrt{\frac{pq}{n}} = \sqrt{\frac{0.5(0.5)}{350}} = 0.0267$$

$$p \pm zs_p = 0.5 \pm 2.054(0.0267) = 0.5 \pm 0.055$$

Se estima con una confianza de 96% que la proporción de trabajadores que cuentan con casa propia está entre 44.5 y 55.5 por ciento.

8.7.1 Determinación del tamaño de muestra para estimar una proporción

Se revisan aquí, al igual que con los procedimientos para estimar una media, 2 casos: cuando se incluye el factor de corrección por población finita y cuando no se le incluye.

8.7.1.1 Cuando no se incluye el factor de corrección por población finita

De la fórmula que resume el procedimiento para estimar la proporción:

$$p \pm z\sigma_p.$$

Si se considera que $z\sigma_p$ es un margen de error alrededor de la proporción, se puede escribir:

$$e = z\sigma_p = z\sqrt{\frac{\pi Q}{n}}$$

Entonces, despejando n de

$$e = z\sqrt{\frac{\pi Q}{n}}$$

se tiene:

$$n = \frac{z^2 \pi Q}{e^2} \qquad \textbf{(8.5)}$$

EJEMPLO 8.10

Se desea tomar una muestra para estimar, con una confianza de 95%, la proporción de artículos defectuosos en un embarque grande y se desea que el error sea, cuando mucho, de 5%. Si se sabe que la proporción de artículos defectuosos en esta clase de embarques fue de 2% en el pasado, determine el tamaño mínimo necesario para la muestra.

$$n = \frac{z^2 \pi Q}{e^2} = \frac{1.96^2(0.02)(0.98)}{0.05^2} = 30.12.$$

El tamaño mínimo de la muestra sería de 31 elementos porque, como se vio para el tamaño de la muestra para estimar una me-

dia, los resultados de estos cálculos siempre se redondean hacia arriba.

Por otra parte, en la determinación del tamaño de la muestra para estimar una proporción, puede darse el caso de que se desconozca totalmente cuál pudiera ser una aproximación razonable de la proporción real y, entonces, sería necesario utilizar un valor de $\pi = 0.5$ porque esto hace que el producto πQ sea el máximo posible: $\pi Q = 0.5(0.5) = 0.25$. Si se utiliza este valor en este ejemplo, se tiene:

$$n = \frac{1.96^2(0.5)(0.5)}{0.05^2} = 384.16$$

que, como puede verse, es un tamaño de muestra mucho mayor que el que se obtuvo con $\pi = 0.02$; por esto siempre resulta más conveniente utilizar alguna estimación preliminar de la proporción de la población porque cualquier valor de π que se aleje de 0.5 dará como resultado un tamaño de muestra menor.

8.7.1.2 Incluyendo el factor de corrección por población finita

En este caso, la expresión que resume el procedimiento para realizar estimaciones de una media es:

$$e = z\sqrt{\frac{pq}{n} \cdot \frac{N-n}{N-1}}$$

Y, despejando n:

$$\frac{e^2}{z^2} = \frac{pq}{n} \cdot \frac{N-n}{N-1} = \frac{pqN - pqn}{nN - n}$$

$$e^2(nN - n) = z^2(pqN - pqn)$$

$$e^2nN - e^2n = z^2pqN - z^2pqn$$

$$e^2nN - e^2n + z^2pqn = z^2pqN$$

$$n(e^2N - e^2 + z^2pq) = z^2pqN$$

$$n = \frac{z^2pqN}{e^2N - e^2 + z^2pq} \tag{8.6}$$

EJEMPLO 8.11

El director de una escuela secundaria desea estimar la proporción de alumnos que cuentan con el esquema completo de vacunas con un margen de error de 3% y un nivel de confianza de 95%. Se sabe que, en el pasado, 70% tenía cubierto el esquema de vacunación. Calcule el tamaño de la muestra mínimo que se requiere, sabiendo que asisten 1 350 alumnos a la escuela.

Solución:

$$n = \frac{z^2pqN}{e^2N - e^2 + z^2pq}$$

$$= \frac{(1.96^2)(0.7)(0.3)(1\ 350)}{(0.03^2)(1\ 350) - 0.03^2 + (1.96^2)(0.7)(0.3)}$$

$$n = \frac{1\ 089.09}{2.02} = 539.15$$

Por lo que el tamaño mínimo necesario de muestra es de 540 estudiantes.

EJEMPLO 8.12

La Secretaría de Educación desea conocer la proporción de estudiantes del nivel medio superior de una zona del país que cuentan con acceso a internet, con un margen de error de 5% y un nivel de confianza de 93%. Por estudios en el pasado se sabe que 57% de los alumnos en este nivel educativo tiene acceso a internet. Calcule el tamaño mínimo de la muestra que se requiere, si los estudiantes del nivel medio superior en la región de interés son 237 520.

Solución:

$$n = \frac{z^2pqN}{e^2N - e^2 + z^2pq} = \frac{(1.815^2)(0.57)(0.43)(237\ 520)}{(0.05^2)(237\ 520) - 0.05^2 + (1.815^2)(0.57)(0.43)}$$

$$n = \frac{191\ 777.1}{594.6} = 322.53$$

Así, el tamaño mínimo de muestra necesario es de 323 estudiantes.

ejercicios 8.7 Estimación de una proporción

1. Una compañía que elabora helados desea estimar con una confianza de 95% la proporción de niños entre 8 y 10 años que prefieren el sabor a chocolate; se tomó una muestra de 150 y se encontró que 87 prefieren el helado sabor a chocolate.
2. Se desea estimar, con una confianza de 99%, la proporción de la población de consumidores de cierta marca de champú que adquiere la presentación de 500 ml. Se tomó una muestra aleatoria de 100 de esos consumidores y se encontró que 37% de ellos compran la presentación de 500 ml.
3. El departamento de limpia de la ciudad de México desea estimar la proporción de hogares que separan la basura con un nivel de confianza de 99%. Se realizaron 500 entrevistas al azar y se encontró que 0.18 de los hogares entrevistados separan su basura.
4. Se desea estimar la proporción de la población de una ciudad grande que utiliza el metro, con un nivel de confianza de 85%. Se encuestaron 600 personas al azar y se encontró que 0.43 utilizan el metro en alguna de sus líneas.
5. Se desea estimar con un nivel de confianza de 94% la proporción de los alumnos que ingresan a licenciatura en una institución universitaria que cuentan con conocimientos de un idioma adicional. Se tomó una muestra aleatoria de 200 alumnos de los cuales 42 dijeron tener conocimientos de otro idioma.

Determinación del tamaño de *n* para estimar *π* sin incluir el factor de corrección por población finita

6. Una empresa productora de dulces desea realizar un estudio de mercado para estimar con un nivel de confianza de 95% la proporción de la población que prefieren sus productos sobre los de la competencia. El error no debe ser mayor a 7%; a raíz de un estudio realizado con anticipación se sabe que la preferencia fue de 39%. Determine el tamaño mínimo para la muestra.
7. Se desea estimar la proporción de artículos devueltos por alguna falla o defecto con un nivel de confianza de 90% y un error no mayor a 6%; en periodos pasados la proporción de devolución fue de 17%. Determine el tamaño mínimo para la muestra.
8. Una cadena de gasolineras quiere estimar, con un nivel de confianza de 97% y un error no mayor a 5%, la proporción de bombas fuera de servicio en los diferentes establecimientos. Determine el tamaño mínimo para la muestra.
9. La dirección de una escuela primaria planea incorporar clases de natación al programa escolar, por lo cual quiere estimar la proporción de alumnos que saben nadar con un nivel de confianza de 99% y un error no mayor a 11%. Determine el tamaño mínimo para la muestra.
10. Se quiere estimar la proporción de alumnos de una escuela preparatoria que cuentan con servicio de internet en casa, con un nivel de confianza de 96% y un error no mayor a 4%, se sabe que en años anteriores la proporción era de 28%. Determine el tamaño mínimo para la muestra.

Incluir el factor de corrección por población finita

11. El departamento de salud preventiva de un centro de salud desea estimar la proporción de enfermos de diabetes que siguen el tratamiento médico de manera correcta con un margen de error de 2.5% y un nivel de confianza de 91%; se conoce que en el pasado esta proporción fue de 1 de cada 3 pacientes. Calcule el tamaño mínimo de la muestra que se requiere si el número de pacientes con diabetes que son atendidos es de 137.
12. En un campus universitario se prestan servicios médicos a los alumnos, la dirección general desea estimar la proporción del estudiantado que hace uso de este servicio con un margen de error de 7.5% y un nivel de confianza de 90%; se sabe que el año pasado 4 de cada 10 estudiantes asistieron a alguna consulta. Calcule el tamaño mínimo de la muestra requerida tomando en cuenta que los estudiantes pertenecientes al campus son 17 350.
13. Se desea estimar la proporción de alumnos en el programa de becas institucionales de una escuela preparatoria que mantienen un promedio de 9 o más y tienen derecho a renovarla, con un margen de error de 4% y un nivel de confianza de 96%; en años anteriores, 58% de los becarios renovaron la beca institucional. Calcule el tamaño mínimo de la muestra requerida, si el padrón total de becarios de la escuela es de 2 720 alumnos.

14. Un agricultor desea estimar la proporción de semillas que se desarrollan en su huerta de tomate con un nivel de confianza de 95% y un margen de error de 10%. En ciclos anteriores de siembra, 72% de las semillas se desarrolló. Calcule el tamaño mínimo de la muestra requerido, si en cada ciclo se plantan 47 000 semillas en la huerta.

15. El departamento de administración escolar desea estimar la proporción de alumnos en el último semestre de la licenciatura de administración que pretenden estudiar alguna maestría, con un nivel de confianza de 97% y un margen de error de 8.5%; anteriormente 31% de los estudiantes expresaron estar interesados en realizar algún tipo de maestría. Calcule el tamaño mínimo de la muestra requerido, si el total de alumnos en el 9o. semestre es de 1 340.

8.8 Otros intervalos de confianza

En los ejemplos siguientes se ilustran otros intervalos de confianza: para la diferencia entre 2 medias poblacionales; para la diferencia entre 2 proporciones poblacionales; para el total de una población a partir de una media y para el total de una población a partir de una proporción.

Es importante no perder de vista que los procedimientos de estimación para la diferencia entre 2 medias y entre 2 proporciones sólo se aplica a los casos en que las 2 muestras son independientes.

8.8.1 Intervalos de confianza para la diferencia entre 2 medias poblacionales

El procedimiento para construir este tipo de intervalos incluye los mismos 3 elementos que ya utilizamos:

1. El valor del estadístico muestral, en este caso, la diferencia entre las medias de las 2 muestras.
2. El valor de z que determina el nivel de confianza.
3. El error estándar del estadístico que, en este caso es el error estándar de la diferencia entre 2 medias y que se calcula como:

$$s_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \tag{8.7}$$

En los siguientes ejemplos se ilustra el procedimiento.

EJEMPLO 8.13

Una empresa desea estimar las horas promedio de trabajo a la semana de las áreas de finanzas y de recursos humanos, para lo cual toma 2 muestras independientes de 130 personas de cada uno de esos departamentos. Del área de finanzas se obtuvo que las horas de trabajo promedio a la semana son 50 con una desviación estándar de 3 horas; en el área de recursos humanos este promedio es de 60 horas con una desviación estándar de 2 horas. Estime la diferencia entre las horas de trabajo de las 2 áreas con un nivel de confianza de 95 por ciento.

Solución:

Finanzas	Recursos Humanos
$n = 130$	$n = 130$
$\bar{X}_1 = 50$	$\bar{X}_2 = 60$
$s = 3$	$s = 2$
$s^2 = 9$	$s^2 = 4$

$$\bar{X}_1 = 50 \text{ y } \bar{X}_2 = 60$$

$$n.c. = 95\%$$

$$z = 1.96$$

$$s_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_1}} = \sqrt{\frac{9}{130} + \frac{4}{130}}$$

$$= \sqrt{0.0692 + 0.0307} = \sqrt{0.0999} = 0.316$$

De lo anterior, el intervalo:

$$(\overline{X}_1 - \overline{X}_2) \pm z s_{\bar{x}_1 - \bar{x}_2} = (60 - 50) \pm 1.96(0.316) = 10 \pm 0.6193$$

Nótese que, al hacer las operaciones, se identifica como muestra 1 a la que tiene el mayor promedio, 60, simplemente para asegurar que la diferencia da un valor positivo, lo cual simplifica los cálculos.

$$10 + 0.6193 = 10.6193$$
$$10 - 0.6193 = 9.3807$$

Por lo que se estima, con un nivel de confianza de 95%, que la diferencia del promedio de horas de trabajo semanales entre las áreas de finanzas y de recursos humanos está entre 9.38 y 10.62 horas.

Ejemplo 8.14

Un banco desea estimar la diferencia entre el promedio del monto depositado en moneda nacional entre los clientes de 2 sucursales, toma una muestra aleatoria de 40 clientes de la sucursal A y otra muestra de igual tamaño de la sucursal B y encuentra que en la primera sucursal se deposita en promedio \$5 000 con una varianza de \$600 y, en la sucursal B, \$3 500 con una varianza de \$700. Construya el intervalo de la diferencia real que existe entre los depósitos de los clientes de las 2 sucursales con un nivel de confianza de 98 por ciento.

Solución:

Sucursal A	Sucursal B
$n = 40$	$n = 40$
$\bar{x} = 5\,000$	$\bar{x} = 3\,500$
$s^2 = 600$	$s^2 = 700$

$n.c. = 98\%$
$z = 2.324$

$$s_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{600}{40} + \frac{700}{40}}$$

$$= \sqrt{15 + 17.5} = \sqrt{32.5} = 5.7$$

De donde, el intervalo

$$(\overline{X}_1 - \overline{X}_2) \pm z s_{\bar{x}_1 - \bar{x}_2} = (5\,000 - 3\,500) \pm 2.324(5.7)$$
$$= 1\,500 \pm 13.246$$
$$1\,500 + 13.281 = 1\,513.246$$
$$1\,500 - 13.281 = 1\,486.75$$

Con un nivel de confianza de 98% se estima que la diferencia real entre los depósitos de los clientes de las 2 sucursales de este banco se encuentra entre \$1 486.75 y \$1 513.25.

8.8.2 Intervalos de confianza para la diferencia entre 2 proporciones poblacionales

Al igual que para todas las estimaciones anteriores, el procedimiento para construir este tipo de intervalos incluye los mismos 3 elementos que ya se utilizan:

1. El valor del estadístico muestral, en este caso, la diferencia entre las proporciones de las 2 muestras.
2. El valor de z que determina el nivel de confianza.
3. El error estándar del estadístico que, en este caso es el error estándar de la diferencia entre 2 proporciones, y que se calcula como:

$$s_{p_1 - p_2} = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}} \qquad \textbf{(8.8)}$$

En los siguientes ejemplos se ilustra el procedimiento.

Ejemplo 8.15

En una delegación política se realizaron encuestas en 2 colonias, con 2 muestras aleatorias independientes de 150 personas cada una para saber su opinión acerca de la construcción de una obra pública; se encontró que en la colonia 1, 90 personas están en favor de la obra; en la colonia 2 hay 75 personas en favor. Construya los límites de confianza para la diferencia entre las proporciones de todos los habitantes de las 2 colonias que están en favor de la obra con un nivel de confianza de 90 por ciento.

Solución:

Colonia 1	Colonia 2
$n = 150$	$n = 150$
$p = 90/150 = 0.60$	$p = 75/150 = 0.50$
$q = 0.40$	$q = 0.50$

Para un nivel de confianza de 90%, $z = 1.64$

$$s_{p_1 - p_2} = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}} = \sqrt{\frac{(0.60)(0.40)}{150} + \frac{(0.50)(0.50)}{150}}$$

$$= \sqrt{0.0016 + 0.0016} = 0.0565$$

$$(p_1 - p_2) \pm z s_{p_1 - p_2} = (0.60 - 0.50) \pm 1.64(0.0565) = 0.10 \pm 0.0926$$

$$0.10 + 0.0926 = 0.1926$$

$$0.10 - 0.0926 = 0.0074$$

Con un nivel de confianza de 90% se estima que la diferencia de la proporción de personas en favor de la construcción de la obra pública entre las 2 colonias es menor a 19.26 por ciento.

EJEMPLO 8.16

Un hospital especializado en cardiología quiere conocer la diferencia entre la eficiencia de 2 tratamientos medicinales y saber si realmente es significativa, por lo que toma 2 muestras independientes, cada una de 200 pacientes; a las personas de la primera muestra les aplica un tratamiento tradicional, mientras que a las de la segunda les aplica uno nuevo. Al cabo de un mes, 170 pacientes de la primera muestra y 110 de la segunda tienen resultados positivos. Construya el intervalo de la diferencia entre las proporciones de la eficiencia de los 2 tratamientos con un nivel de confianza de 94 por ciento.

Solución:

Tratamiento tradicional	Tratamiento nuevo
$n = 200$	$n = 200$
$p = 0.85$	$p = 0.55$
$q = 0.15$	$q = 0.45$

n.c. = 94%
$z = 1.88$

$$s_{p_1 - p_2} = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}} = \sqrt{\frac{(0.85)(0.15)}{200} + \frac{(0.55)(0.45)}{200}}$$

$$= \sqrt{0.00063 + 0.0012} = 0.0427$$

Y, el intervalo

$$(p_1 - p_2) \pm z s_{p1 - p2} = (0.85 - 0.55) \pm 1.88(0.0427) = 0.30 \pm 0.0802$$

$$0.30 + 0.0802 = 0.3802$$

$$0.30 - 0.0802 = 0.2198$$

Se estima con un nivel de confianza de 94% que la diferencia de la proporción de pacientes que presentaron resultados positivos en los 2 tratamientos está entre 21.98 y 38.02 por ciento.

8.8.3 Intervalos de confianza para el total de una población a partir de una media

Se ilustra en los siguientes ejemplos el procedimiento para estimar un total poblacional cuando se conoce la estimación de la media y el número de elementos de la población.

EJEMPLO 8.17

En el ejemplo 8.1 se construyó un intervalo de confianza para las ventas mensuales de un nuevo alimento para perros en 36 tiendas de autoservicio y se estimó, con un nivel de confianza de 95%, que el promedio real de ventas para este nuevo alimento para perros era de entre \$11 738.67 y \$12 261.33; si ahora se añade el dato de que la cadena de tiendas tiene un total de 300 establecimientos, se puede construir un intervalo de confianza para el *total* de las ventas simplemente multiplicando esos valores promedio por el número total de tiendas:

$$11\,738.67(300) = \$3\,521\,601$$

$$12\,261.33(300) = \$3\,678\,399.$$

Entonces, se estima, con una confianza de 95%, que el total de ventas mensuales de todas las tiendas de la cadena está entre \$3 521 601 y \$3 678 399.

EJEMPLO 8.18

En el ejemplo 8.2 se construyó un intervalo de confianza de 90% para el promedio del ingreso diario de 50 000 peones de la construcción que laboran en el Distrito Federal con una muestra aleatoria de 400 de ellos, y se estimó que su ingreso promedio diario

estaba entre \$73.13 y \$74.87; con estos datos, y sabiendo que son 50 000 el total de trabajadores, el intervalo de confianza para este total es:

$$73.13(50\,000) = \$3\,656\,500$$

$$74.87(50\,000) = \$3\,743\,500.$$

De lo anterior, se estima con un nivel de confianza de 90%, que el total de los salarios diarios de los 50 000 peones de la construcción que laboran en el Distrito Federal está entre \$3 656 500 y \$3 743 500.

8.8.4 Intervalos de confianza para el total de una población a partir de una proporción

Se ven en seguida ejemplos que muestran el procedimiento para estimar un total cuando se conoce la estimación de la proporción y el número de elementos de la población.

Ejemplo 8.19

En el ejemplo 8.8 se estimó, con una confianza de 99% y con base en una muestra de 200 alumnos, que la proporción de alumnos de una universidad que acuden a sus instalaciones en su propio automóvil está entre 6.5 y 18.5%; si se sabe que el total de estudiantes de esa universidad es de 2 000, el intervalo de confianza para el total de estudiantes que acuden a la escuela en automóvil es de:

$$2\,000(0.065) = 130$$

$$2\,000(0.185) = 370.$$

De donde, se estima, con un nivel de confianza de 99%, que el total de estudiantes que acuden a la universidad en su propio automóvil está entre 130 y 370.

Ejemplo 8.20

En el ejemplo 8.9 se construyó un intervalo de confianza de 96% para la proporción de los trabajadores de una empresa que cuentan con casa propia a partir de una muestra de 350 empleados, se estimó que esa proporción estaba entre 45 y 55%; si en esa empresa laboran 450 empleados, se puede estimar, con esa misma confianza de 96%, que el total de trabajadores que tienen casa propia está entre 157.5 y 192.5 (350×0.45 y 350×0.55, respectivamente).

Ejercicios 8.8 Otros intervalos de confianza

Estimación de la diferencia entre 2 medias

1. Una dependencia del gobierno desea estimar la diferencia entre el nivel de vida de 12 zonas del país, obtiene una muestra aleatoria de 500 personas del sur del país con un salario promedio de \$3 000 con una varianza de \$400 y otra muestra del mismo tamaño, pero de la zona norte, con un salario promedio de \$8 000 con una varianza de \$700. Construya el intervalo de la diferencia de los salarios promedio entre estas 2 zonas con un nivel de confianza de 95 por ciento.
2. Dos personas del área de recursos humanos tienen como trabajo reunir cierta cantidad de personas que cumplan con determinadas características para ser contratados como vendedores, la obtención de estos cargos requiere de un examen de habilidad verbal que tiene una escala de evaluación que va de 0 a 3 000; el gerente de esta área quiere estimar la diferencia del promedio de calificación obtenida entre los 2 grupos de candidatos correspondientes a cada uno de los trabajadores de recursos humanos, por lo que toma una muestra aleatoria de 40 candidatos de cada grupo y encuentra que en el primero la calificación promedio fue de 2 600 con una desviación estándar de 150 y, en el segundo grupo, 2 200 con una desviación estándar de 180. Estime con un nivel de confianza de 99% la diferencia entre las 2 medias y explicar qué podríamos deducir de este resultado.
3. Una compañía especializada en encontrar estudiantes con cierto perfil de acuerdo con las vacantes de las empresas con las que trabaja desea estimar la diferencia del promedio de tiempo que permanecen en la empresa una vez contratados, entre los que recibieron capacitación y aquellos que no; de acuerdo con los registros de seguimiento, obtiene una muestra aleatoria de 85 estudiantes contratados capacitados y 85 no capacitados y encuentra que, del primer grupo, el promedio de tiempo de permanencia en la empresa es de 7.5 años con una desviación estándar de 1.3 y del grupo de los que no recibieron capa-

citación el tiempo promedio es de 4 años con una desviación estándar de 0.8.

Con un nivel de confianza de 90%, encuentre la diferencia de promedios de tiempo de permanencia en la empresa entre los estudiantes que recibieron capacitación y los que no.

4. Una empresa especializada en limpieza de maquinaria fabril ofrece 2 tipos de servicio: básico y avanzado, por lo que quiere saber la diferencia del promedio de tiempo en que los clientes que utilizan uno de los 2 servicios los contactan para llevar a cabo la siguiente limpieza; para esto, toma una muestra aleatoria de 32 clientes de cada grupo y encuentra que del primero tardan en promedio 6 meses para solicitar la siguiente limpieza con una desviación estándar de 0.7 y del segundo este promedio es de 10 meses con una desviación estándar de 0.95. Estime la diferencia de tiempo promedio en que los clientes programan la próxima cita entre estos 2 grupos con un nivel de confianza de 94 por ciento.

5. Una tienda departamental desea estimar la diferencia del promedio de las ventas realizadas entre el departamento de caballeros y el de damas, por lo que toma una muestra aleatoria de 150 ventas de cada departamento correspondientes a una semana y encuentra que las del departamento de damas tienen una valor promedio de $430 con una desviación estándar de $25 mientras que, en el de los hombres, son de $270 con una desviación estándar de $17. Encuentre la diferencia de los promedios de las ventas entre los 2 departamentos con un nivel de confianza de 99 por ciento.

Estimación de la diferencia entre 2 proporciones

6. Una empresa de servicios financieros tiene una propuesta acerca de un nuevo proyecto de inversión, para lo cual es necesario la aprobación de la mayoría de los accionistas; el consejo de dicha empresa se integra por 2 grupos: el correspondiente a los socios nacionales y el segundo a los extranjeros. Para llevar a cabo el proyecto es muy importante que la diferencia de proporciones de accionistas que aprueben el proyecto no sea grande, aunque los socios mexicanos conformen la mayor parte del consejo. Para estimar esto, se toma una muestra aleatoria de 70 accionistas mexicanos, de los cuales 57 están de acuerdo con el proyecto, y una de muestra aleatoria de 40 extranjeros, de los cuales 23 desearían llevar a cabo el proyecto de inversión. Con un nivel de confianza de 90%, estime la diferencia de proporciones entre los accionistas nacionales y extranjeros que están de acuerdo con el proyecto y con base en esto, decidir si el resultado es representativo.

7. Una empresa industrial de artículos deportivos divide su producción en 2 áreas importantes: una fabrica zapatos para la práctica de diferentes deportes y otra ropa; los jefes de operación de las 2 áreas desean estimar la diferencia entre las proporciones de artículos que se venden. De una muestra aleatoria de 800 zapatos producidos, 679 son vendidos la misma semana, mientras que en el área de ropa se venden 260 artículos de una muestra aleatoria de 400 fabricados. Estime, con un nivel de confianza de 94%, la diferencia entre las proporciones de artículos que se venden semanalmente entre estas 2 áreas para que los jefes de operación puedan tomar decisiones con base en el resultado.

8. Un país europeo desea lanzar un programa de intercambios académicos para estudiantes de nivel superior con México, por lo que desea estimar la diferencia de proporciones entre alumnos de ese país y alumnos de México que muestran interés en llevar a cabo un intercambio a lo largo de su carrera universitaria. De una muestra aleatoria de 1 500 estudiantes mexicanos, 680 de ellos dijeron estar interesados en hacer el intercambio al lugar propuesto, mientras que de la muestra aleatoria del mismo tamaño correspondiente al país europeo, a 850 estudiantes les interesaría estudiar en México durante la carrera. Estime la diferencia entre las proporciones de estos 2 grupos con un nivel de confianza de 98 por ciento.

9. El gerente de ventas de una distribuidora de sistemas de información computarizados para el registro de nómina desea estimar la diferencia que existe entre las proporciones de empresas públicas y privadas que utilizan sus sistemas, por lo que, a través de encuestas aplicadas a una muestra aleatoria de 70 empresas de cada grupo, encuentra que 37 públicas y 52 privadas ocupan estas herramientas. Estime, con un nivel de confianza de 99%, la diferencia de proporciones entre empresas públicas y privadas que los utilizan para que, con este dato, el gerente sepa si vale la pena concentrarse en algún mercado en particular.

10. Los encargados del área de mercadotecnia de una empresa desean estimar la diferencia de proporciones entre clientes potenciales hombres y mujeres, por lo que dan a conocer su producto a 2 muestras aleatorias de 80 hombres y 80 mujeres; de la primera, 23 de ellos dijeron estar de acuerdo en comprar el producto y, de la segunda, 55 dicen lo mismo. Estime la diferencia entre las proporciones de hombres y mujeres que adquirirían el producto con un nivel de confianza de 97 por ciento.

Estimación de un total a partir de una media

11. Se tomó una muestra de 35 empleados de una empresa, se estimó, con 95% de nivel de confianza, que el promedio real de los salarios se encuentra entre $131.01 y $134.99, sabiendo que en total laboran 500 empleados en la empresa, construya un intervalo de confianza para el total de salarios.

12. Una cadena de comida rápida registró el número de clientes que recibió durante el día en una muestra aleatoria de 43 de los restaurantes que tiene en el país, se estimó, con nivel de confianza de 90%, que el promedio real de clientes se encuentra entre 101.23 y 112.77. Si

la cadena cuenta con 730 restaurantes en todo el país, construya un intervalo de confianza para el total de clientes que se reciben diariamente.

13. Se tomó una muestra aleatoria de 400 llantas producidas en una planta, y se encontró que la vida útil promedio era de 43 000 kilómetros con una desviación estándar de 1 300 kilómetros; con un nivel de confianza de 85%, se estimó que el promedio real de la vida útil de la llanta se encuentra entre 42 067.25 y 43 932.75 kilómetros. Construya un intervalo de confianza para el total de vida útil en kilómetros de las llantas, sabiendo que la producción total es de 200 000 piezas.

14. En 33 oficinas postales elegidas al azar se recibieron en promedio 1 357 cartas durante un día específico con una desviación estándar de 226, se estimó, con un nivel de confianza de 92%, que el promedio de cartas recibidas al día se encuentra entre 1 287.9 y 1 426.1. Construya un intervalo de confianza para el total de oficinas postales sabiendo que existen 5 200.

15. De los 15 000 alumnos que asisten a una institución universitaria, se preguntó a una muestra aleatoria de 120 de ellos cuánto tiempo les tomaba llegar a la escuela, se hizo una estimación con nivel de confianza de 90% y se encontró que en promedio les toma llegar a la escuela entre 26.34 y 29.82 minutos. Construya un intervalo de confianza para el total del tiempo de traslado de todos los estudiantes.

Estimación de un total a partir de una proporción

16. Una compañía que elabora helados desea estimar con una confianza de 95% la proporción de niños entre 8 y 10 años que prefieren el sabor a chocolate; se tomó una muestra de 150 niños entre esas edades y se encontró que la proporción de niños que prefieren el helado de chocolate está entre 50 y 66%. Construya un intervalo de confianza para el total de niños que prefieren el helado de chocolate, sabiendo que el mercado de esa empresa consta de 565 000 niños de entre 8 y 10 años.

17. Se desea estimar, con una confianza de 99%, la proporción de la población de consumidores de cierta marca de champú que adquiere la presentación de 500 ml. Se tomó una muestra aleatoria de 100 de esos consumidores y se encontró que la proporción de consumidores que adquieren esa presentación está entre 25 y 49%. Construya un intervalo de confianza para el total de consumidores del champú que compran la presentación de 500 ml sabiendo que existen 234 600 personas que consumen la marca.

18. El departamento de limpia de la ciudad de México desea saber la proporción de hogares que separan la basura con un nivel de confianza de 99%; se realizaron 500 entrevistas al azar y se estimó que la proporción de hogares que sí lo hacen se encuentra ente 14 y 22%. Construya un intervalo de confianza para el total de hogares sabiendo que existen 2 180 240 de hogares en la ciudad de México.

19. Se desea estimar, con un nivel de confianza de 85%, la proporción de la población de una ciudad grande que utiliza el metro, se encuestaron a 600 personas al azar y se encontró que la proporción que usa el metro está entre 40 y 46%. Construya un intervalo de confianza para el total de personas que utilizan el metro sabiendo que en la ciudad existen 3 250 320 habitantes.

20. Se desea estimar, con un nivel de confianza de 94%, la proporción de los alumnos que ingresan a licenciatura en una institución universitaria contando con conocimientos de un idioma adicional; se tomó una muestra aleatoria de 200 alumnos y se encontró que la proporción que tiene conocimientos de otro idioma está entre 16 y 26%. Construya un intervalo de confianza para el total de estudiantes sabiendo que ingresaron 1 732.

8.9 Resumen

En este capítulo se revisaron los procedimientos estadísticos que se aplican en la estimación de parámetros, que es una de las 2 grandes técnicas de la inferencia estadística (pruebas de hipótesis es la otra).

Se comenzó por revisar la diferencia entre estimaciones por punto y estimaciones por intervalo y se vio que estas últimas son más convenientes que las primeras porque a las estimaciones por intervalo se les pueden asociar criterios de probabilidad (confiabilidad), cosa que no es posible con las estimaciones por punto.

Se analizaron también las 4 principales propiedades deseables en los estimadores: la ausencia de sesgo, la consistencia, la eficiencia y la suficiencia.

El resto del capítulo se ocupó de analizar y presentar ejemplos de los procedimientos para realizar las estimaciones de parámetros más utilizadas:

- De una media (secciones 8.4 a 8.6).
- De una proporción (sección 8.8).
- De la diferencia entre 2 medias (sección 8.8.1).
- De la diferencia entre 2 proporciones (sección 8.8.2).
- De un total a partir de una media (sección 8.17).
- De un total a partir de una proporción (sección 8.8.4).

Desde los primeros ejemplos de estimaciones de parámetros se hizo hincapié en la conveniencia de tener presentes los 3 elementos de una estimación:

1. El estadístico muestral alrededor del cual se construye el intervalo.
2. El nivel de confianza especificado que se utiliza para determinar el valor de z, la desviación estándar de la distribución normal estandarizada, que es la que permite medir las pro-

babilidades de acuerdo con esta distribución (o el valor de t, para el caso de muestras pequeñas).

3. El error estándar del estimador que puede ser, entonces, el error estándar de la media, de la proporción, de la diferencia entre 2 medias o de la diferencia entre 2 proporciones, según sea el caso.

Respecto a la estimación de intervalo de una media se hizo notar la diferencia entre llevarla a cabo con muestras grandes y con muestras pequeñas y se vio que, con muestras grandes, siempre es posible utilizar la z normal como estadístico teórico, en tanto que, cuando se trata de muestras pequeñas, sólo se puede utilizar esta z cuando se conoce la desviación estándar de la población y se puede asumir que la variable de interés se distribuye de manera normal en la población. Además se revisó que, cuando ese tamaño de muestra es pequeño, la variable se distribuye de manera normal y se desconoce la desviación estándar de la población, entonces se puede utilizar la desviación estándar de la muestra como estimador de esa σ y como estadístico teórico la t de Student.

Asimismo, se estudió el procedimiento para determinar el tamaño de muestra mínimo necesario para estimar tanto una media como una proporción y con 2 casos en ambas estimaciones: utilizando y no utilizando el factor de corrección por población finita.

8.10 EXCEL Uso de Excel para construir intervalos

Existe solamente una función en Excel para construir intervalos de confianza:

INTERVALO.CONFIANZA(alfa;desv_estándar;tamaño)

la cual sólo se puede aplicar para la estimación por intervalo de una media poblacional. Alfa es el valor que se obtiene al restarle a 1 el nivel de confianza; "desv_estándar" es la desviación estándar de la población cuando se sabe su valor o la desviación estándar de la muestra cuando la primera es desconocida y, finalmente, "tamaño" es el tamaño de la muestra. Para ilustrar su utilización, en seguida se repasan algunos de los ejemplos resueltos en el texto.

En el ejemplo 8.1 se estimó, con una confianza de 95%, el promedio real de ventas de un nuevo alimento para perros utilizando el promedio de las ventas de 36 tiendas durante un mes, y se encontró un promedio de \$12 000 de ventas por tienda, con desviación estándar de \$800. Se usa la función de Excel de la siguiente manera:

=INTERVALO.CONFIANZA(0.05,800,36)

Con lo que se obtiene 261.3285, que es el valor que hay que sumar y restar a la media de la muestra para obtener el mismo intervalo que se construyó antes.

$$\overline{X} \pm z\sigma_{\overline{X}} = \overline{X} \pm zs_{\overline{X}} = 12\,000 \pm 261.33$$

Obsérvese que entonces que se utilizó un $\alpha = 1 - 0.95 = 0.05$. Este valor se relaciona con el tema de "nivel de significación" que se tratará con detalle en el capítulo siguiente sobre pruebas de hipótesis. De momento, basta con saber que se calcula como $\alpha = 1 -$ nivel de confianza. Con este nivel de confianza, Excel determina el valor correspondiente de z y hace las operaciones necesarias para calcular ese valor que se suma y se resta de la media para obtener el intervalo de confianza.

En el ejemplo 8.5, que ilustra el caso de estimación con muestras pequeñas, trata de una empresa que fabrica focos para iluminación de escenarios y que tienen una desviación estándar de 40 horas; se tomó una muestra aleatoria de 25 focos y se encontró que su vida útil promedio era de 835 horas. El intervalo de confianza de 95% que se construyó fue:

$$\overline{X} \pm z\sigma_{\overline{X}} = 835 \pm 1.96(8) = 835 \pm 15.68$$

Utilizando ahora Excel:

=INTERVALO.CONFIANZA(0.05,40,25)

Se obtiene 15.67971, que es el mismo valor calculado antes.

En resumen, puede resultar útil esta función de Excel, sobre todo cuando sea necesario llevar a cabo muchas estimaciones, lo que se debe tener presente en todo momento es que esta función sólo es aplicable cuando el estadístico teórico es la z, es decir, cuando se hacen estimaciones de la media con muestras

grandes o, en un segundo caso, cuando se hacen estimaciones de la media con muestras pequeñas, la variable se distribuye de manera normal y se conoce la desviación estándar de la población, que son los 2 casos que, según se vio en este capítulo, son en los que se utiliza la distribución normal para construir el intervalo.

8.11 Fórmulas del capítulo

8.4 Estimación de una medida con muestras grandes

Procedimiento para construir una estimación por intervalo para una media aritmética:

$$\overline{X} \pm z\sigma_{\overline{X}} \tag{8.1}$$

8.4.4 Determinación del tamaño de muestra necesario para estimar una media

Tamaño de muestra para estimar una media, sin utilizar el factor de corrección por población finita:

$$n = \left(\frac{z\sigma}{e}\right)^2 \tag{8.2}$$

Tamaño de muestra para estimar una media cuando se utiliza el factor de corrección por población finita:

$$n = \frac{Nz^2\sigma^2}{e^2(N-1) + z^2\sigma^2} \tag{8.3}$$

8.7 Estimación de una proporción

Procedimiento para construir una estimación por intervalo para una proporción:

$$p \pm z\sigma_p \tag{8.4}$$

8.7.1 Determinación del tamaño de muestra para estimar una proporción

Tamaño de muestra para estimar una proporción sin utilizar el factor de corrección por población finita:

$$n = \frac{z^2\pi Q}{e^2} \tag{8.5}$$

Tamaño de muestra para estimar una proporción cuando se utiliza el factor de corrección por población finita:

$$n = \frac{z^2pqN}{e^2N - e^2 + z^2pq} \tag{8.6}$$

8.8.1 Intervalos de confianza para la diferencia entre 2 medias poblacionales

El error estándar de la diferencia entre 2 medias:

$$s_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \tag{8.7}$$

8.8.2 Intervalos de confianza para la diferencia entre 2 porporciones poblacionales

El error estándar de la diferencia entre 2 proporciones:

$$s_{p_1 - p_2} = \sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}} \tag{8.8}$$

8.12 Ejercicios adicionales

8.4 Estimación de una media con muestras grandes

1. En una muestra de 30 librerías elegidas al azar en una ciudad grande se vendieron en promedio 843 libros por semana, con una desviación estándar de 32. Estime el intervalo para el número promedio de libros vendidos semanalmente en todas las librerías de la ciudad con un nivel de confianza de:

   a) 75%.
   b) 80%.

2. Un banco tomó una muestra aleatoria de 100 tarjetahabientes que presentan retraso en sus pagos; en promedio, cada cliente debe $4 720 con una desviación estándar de $558. Estime el intervalo para la deuda promedio de todos los clientes con atraso en sus pagos con un nivel de confianza de 90 por ciento.
3. Una empresa de servicios telefónicos quiere incrementar sus llamadas de larga distancia internacional implementando una tarifa más baja para sus clientes; dicho proyecto sólo es factible si el promedio de todas las llamadas supera los 4 min, por lo que el área de proyectos toma una muestra aleatoria de 50 llamadas internacionales y encuentra que tienen una duración promedio de 3.7 min con una desviación estándar de 0.6 minutos. Construya el intervalo correspondiente al tiempo promedio de duración de todas las llamadas a larga distancia internacional de acuerdo con los datos muestrales y con un nivel de confianza de 90% y explique por qué la empresa debe rechazar o aceptar la propuesta.
4. La directora de una escuela primaria desea saber el promedio académico general de sus alumnos, toma una muestra aleatoria de 120 alumnos y observa que el promedio general es de 8.2 con una desviación estándar de 0.2.

   Construya el intervalo del promedio académico de todos los alumnos de la primaria con un nivel de confianza de 98 por ciento.
5. Se tomó una muestra de 80 cajas de focos que contienen 50 focos cada una, se contaron las piezas con algún tipo de defecto en cada caja. Se presentan los resultados a continuación. Estime el intervalo con un nivel de confianza de 90% para el total de cajas producidas.

Piezas defectuosas *X*	Número de cajas *f*
0 a menos de 3	3
3 a menos de 6	5
6 a menos de 9	18
9 a menos de 12	24
12 a menos de 15	15
15 a menos de 18	9
18 a menos de 21	6
Total	80

6. Del total de llamadas que realizaron los usuarios de una compañía de telefonía móvil durante 30 días, se tomó una muestra de 300 llamadas y se registró su duración. Los resultados se presentan en la siguiente tabla. Estime el intervalo con un nivel de confianza de 95% para el total de llamadas realizadas.

Duración *X*	Número de llamadas *f*
1 a menos de 4	8
4 a menos de 7	11
7 a menos de 10	25
10 a menos de 13	109
13 a menos de 16	94
16 a menos de 19	34
19 a menos de 22	13
22 a menos de 25	6
Total	300

8.4.4.1 Cuando no se incluye el factor de corrección por población finita

7. Se quiere estimar, con una confianza de 90%, el número de libros con algún tipo de daño del acervo de la biblioteca de una escuela preparatoria, se sabe por estudios anteriores que la desviación estándar del número de libros con alguna mutilación es de 19. Determine el tamaño mínimo de la muestra requerida para estimar el promedio de libros con algún daño en toda la biblioteca con un margen de error inferior a 2.

8. Se sabe que la desviación estándar del número de horas que pasan frente a una computadora los trabajadores de cierta empresa es de 3.5, se quiere saber el número promedio que pasan los trabajadores de toda la organización trabajando frente a una computadora. Determine el tamaño mínimo de la muestra con un nivel de confianza de 93% y con un margen de error inferior a 0.25.

9. La desviación estándar del diámetro de las varillas fabricadas por una máquina es de 7 milímetros, el gerente de producción quiere saber el diámetro promedio de las varillas producidas. Determine el tamaño mínimo de la muestra con un nivel de confianza de 85% y un margen de error menor a 1.5 milímetros.

10. En una tienda departamental se regalan cupones de precio especial por cada $100 de compra, el gerente de ventas quiere estimar cuál es la compra promedio de los clientes que usan estos cupones. Determine el tamaño mínimo de la muestra, con un nivel de confianza de 99% y un margen de error menor a 10, si sabe que la desviación estándar de estas compras es de $47.

8.4.4.2 Cuando sí se incluye el factor de corrección por población finita

11. Se desea conocer el promedio del monto que tienen las ventas que se hacen a través del portal en internet de una distribuidora de electrodomésticos, con un nivel de confianza de 98% y un margen de error de 2; se conoce que la desviación estándar es de 41. Calcule el tamaño de muestra mínimo que se requiere sabiendo que el número de ventas a través del portal es de 1 230.

12. Por cuestiones de control, la semana pasada se evaluaron los conocimientos de todo el personal operativo de una fábrica armadora mediante un examen con escala de evaluación del 1 al 10, el director general de la empresa desea estimar el promedio de la calificación obtenida por los 8 500 empleados con un nivel de confianza de 95%, sin que el error de la estimación sea superior a 0.2 y considerando que la varianza de la población es de 0.8. Calcule el tamaño de la muestra necesario para poder obtener dicha estimación.

13. El administrador de un lujoso club deportivo de la ciudad de México desea estimar el promedio de visitas que realizan los 917 socios del club durante un mes, el error de dicha estimación no puede ser superior a 0.5 y, además, se requiere tener un nivel de confianza de 97%. Calcule cuántos socios deben considerarse para poder realizar la estimación considerando que la varianza es de 6.25.

14. Un distribuidor de productos electrónicos desea estimar el promedio de meses de garantía que tienen los 175 productos con los que comercia, con un nivel de confianza de 96%; de acuerdo con las estimaciones de otro distribuidor de productos muy parecidos, contempla una varianza de 1.44 y desea que la estimación esté dentro del 0.40 del promedio real. ¿Cuántos productos debe tomar en cuenta para poder realizar la estimación?

15. Una agencia de publicidad desea estimar el promedio de gastos que realizaron los 157 clientes a los que atendió el año pasado esperando un nivel de confianza de 93%; de acuerdo con una prueba piloto se sabe que la varianza es de 40 y espera que la estimación se encuentre dentro de 2% alrededor del valor real. Determine el tamaño de la muestra necesario para determinar esta estimación.

8.6 Estimación de una media con muestras pequeñas

8.6.1 Se conoce la desviación estándar de la población

16. En un campus universitario se organizan jornadas mensuales de conferencias y se sabe que la desviación estándar del número de personas que asisten a cada ponencia es de 12, se seleccionaron 17 conferencias al azar y se encontró que en promedio asisten 108 personas. Construya un intervalo

de confianza de 94% para el número de asistentes a las conferencias.

17. Un producto se comercializa en diferentes establecimientos comerciales y la desviación estándar de su precio es de $4.50, se tomó el precio de 12 establecimientos diferentes elegidos al azar y se encontró que, en promedio, el producto cuesta $124.30. Construya un intervalo del confianza de 91% para el precio promedio del producto.

18. Una granja familiar comercializa los huevos que obtiene de sus gallinas y el número de huevos que producen al año tiene una desviación de 20, se tomó una muestra de 10 años y se encontró que, en promedio, se obtienen 1 300 huevos al año. Construya un intervalo de confianza de 97% para la producción promedio anual de huevos.

19. La edad de los trabajadores de una empresa tiene una desviación estándar de 8 años, se le preguntó a 11 trabajadores y se encontró que su edad promedio era de 33 años. Construya un intervalo de confianza de 90% para la edad promedio de los trabajadores de la empresa.

20. El número de teléfonos celulares que ingresan al mes a servicio técnico por alguna falla tiene una desviación estándar de 18, se registraron los ingresos de 9 meses elegidos al azar y se encontró que el promedio de celulares que ingresan a servicio es de 54. Construya un intervalo de confianza de 96% para el número promedio de teléfonos celulares que ingresan mensualmente a servicio.

8.6.2 No se conoce la desviación estándar de la población pero se sabe que la variable se distribuye de manera normal

21. Una tienda de muebles quiere saber el promedio de unidades vendidas diariamente en el último año para compararlo con el registro de las ventas diarias de años anteriores, por lo que toma una muestra aleatoria de las ventas de 7 días.

Día	1	2	3	4	5	6	7
X Unidades vendidas	82	95	100	87	92	76	80

a) Calcule la media puntual correspondiente a las unidades vendidas diariamente durante ese mes.

b) Estime la desviación estándar de la población.

22. El gerente de ventas de una empresa que publica el periódico local de una pequeña ciudad desea enfocarse el próximo trimestre a obtener mayores ventas por mayoreo, ya que esto representa más utilidades y menos gastos para la empresa; además, de esta forma podría aclararse si los representantes de ventas dan los resultados esperados, con una muestra aleatoria de 22 semanas, el gerente encuentra que los vendedores realizan cada semana un promedio de 15 visitas a clientes de este tipo con una desviación estándar de 2.5 visitas. Asumiendo que las ventas se distribuyen de manera normal y con un nivel de confianza de 90%, estime el promedio de visitas semanales por vendedor.

23. Una ensambladora de procesadores para computadoras tiene una propuesta para armar cada uno de sus productos en 30 minutos, por lo que desea evaluar el tiempo promedio actual que se le invierte al ensamblaje de su producción; con una muestra aleatoria de 17 procesadores, se observa que en promedio toma 38 minutos el ensamblado de cada uno de ellos con una desviación estándar de 1.9 min, si el tiempo de ensamblado se distribuye normalmente, estime, con un nivel de confianza de 94%, el promedio del tiempo requerido para ensamblar cada procesador y decidir si es conveniente cambiar el método de ensamblado por el nuevo que se propone.

24. Una empresa de asesoría financiera desea saber el tiempo promedio en que sus clientes con inversiones mayores a $500 000 recuperan su dinero; con una muestra aleatoria de 12 clientes que tardan en promedio 3 años para recuperar su inversión y con una desviación estándar de 0.5 años, estime el promedio del tiempo en que todos los clientes con inversiones superiores a $500 000 recuperan su dinero con un nivel de confianza de 99%. Se asume que el tiempo de recuperación de la inversión se distribuye de manera normal.

25. Los accionistas de una empresa de manufactura quieren invertir por primera vez un porcentaje de las utilidades de los últimos 5 años en acciones de otras empresas mexicanas, esperando obtener dividendos de ellas; los socios toman una muestra aleatoria de 15 empresas nacionales que en promedio registran dividendos de $100 000 con una desviación estándar de $5 000. Estime, con un nivel de confianza de 95%, el promedio de dividendos que obtendrían si invirtieran en una empresa mexicana si se considera que los dividendos se distribuyen normalmente.

26. Una muestra de 29 trabajadores de una productora de cartón armó cajas durante una hora, los resultados se muestran en la siguiente tabla; si siguen una distribución aproximadamente normal, estime el intervalo con un nivel de confianza de 80% para el total de trabajadores.

Número de cajas *X*	Número de trabajadores *f*
15 a menos de 20	3
20 a menos de 25	5
25 a menos de 30	10
30 a menos de 35	7
35 a menos de 40	4
Total	29

8.7 Estimación de una proporción

27. Una aerolínea desea saber la proporción de la población que realizó algún viaje en avión durante los últimos 6 meses con un nivel de confianza de 98%; se encuestaron a 250 personas, de las cuales 0.37% realizó algún viaje en avión.

28. Se quiere saber la proporción de la población que lee el periódico diariamente con un nivel de confianza de 95%, se tomó una muestra de 400 personas, de las cuales 128 dijeron leer el periódico diariamente. Haga la estimación.

29. Se quiere saber con un nivel de confianza de 99% la proporción de la población de universitarios que desearían obtener una beca para estudiar en el extranjero, se tomó una muestra de 320, de los cuales, 41% desearían obtener una beca para estudiar en el extranjero. Estime el intervalo.

30. En una escuela primaria se quiere conocer la proporción de alumnos que usan lentes con un nivel de confianza de 90%; se seleccionó una muestra de 80 alumnos, de los cuales 40 utilizan lentes. ¿Cuál es la estimación por intervalo?

31. Se quiere estimar con un nivel de confianza de 99% la proporción de habitantes de una población rural que cuentan con algún tipo de seguridad social, se tomó una muestra de 300 habitantes, de los cuales 150 cuentan con seguridad social. Haga la estimación correspondiente por intervalo.

8.7.1.1 Determinación del tamaño de muestra para estimar una proporción cuando no se incluye el factor de corrección por población finita

32. Un hospital quiere saber, con un nivel de confianza de 85% y un error no mayor a 8%, la proporción de pacientes que son ingresados y requieren tratamiento quirúrgico; se sabe por estudios anteriores que 9% de los pacientes que ingresan necesitan tratamiento quirúrgico. Determine el tamaño mínimo para la muestra.

33. Se desea saber la proporción de llamadas que entran al centro de atención al cliente para realizar alguna queja, con un nivel de confianza de 95% y un error no mayor a 5%; se sabe que en el pasado la proporción fue de 37%. Determine el tamaño mínimo para la muestra.

34. Una empresa que ofrece servicios de televisión por cable quiere conocer con un nivel de confianza de 93% y un error no mayor a 9%, la proporción de hogares que adquieren el paquete completo de televisión, en el pasado esta proporción fue de 23%. Determine el tamaño mínimo para la muestra.

35. Se quiere conocer la proporción de personas que viajaran en los autobuses de la línea MexaBus con destino a Acapulco en las próximas vacaciones de Semana Santa, con un nivel de 99% y un error no mayor a 10%; se sabe que en el pasado la proporción fue de 14%. Determine el tamaño mínimo para la muestra.

36. Se quiere saber la proporción de la población de una pequeña comunidad rural que terminó la educación primaria con un nivel de confianza de 90% y un error no mayor a 7%. Determine el tamaño mínimo para la muestra.

8.7.1.2 Incluyendo el factor de corrección por población finita

37. Se quiere conocer la proporción de niños de entre 6 y 10 años de edad de una población determinada que tiene caries, con un nivel de confianza de 99% y un margen de error de 5%; en estudios anteriores se encontró que 18% de los niños de estas edades en la población tienen caries. Calcule el tamaño mínimo requerido de la muestra si en la población existen 3 418 niños de entre 6 y 10 años de edad.

38. Se desea conocer la proporción de personas que entran diariamente a una tienda departamental y realizan una compra, con un nivel de confianza de 94% y un margen de error de 2%, se sabe que 3 de cada 10 clientes realizan una compra al ingresar a la tienda. Calcule el tamaño mínimo requerido de la muestra si al día entran 724 personas.

39. Una agencia de investigación de mercado desea conocer la proporción de hogares en los que se ve la televisión en cierto horario en una zona geográfica determinada, con un nivel de confianza de 92% y un margen de error de 6.5%; en estudios anteriores se encontró que, en el horario seleccionado, 61% de los hogares ven televisión. Calcule el tamaño mínimo requerido de la muestra si existen 14 357 hogares en la zona a estudiar.

40. La gerencia de recursos humanos desea saber, con un nivel de confianza de 95% y un margen de error de 2%, la proporción de ingresos que los trabajadores de la empresa destinan al ahorro, se conoce que en años anteriores el trabajador destinaba 23% de sus ingresos al ahorro. Calcule el tamaño mínimo requerido de la muestra si en la empresa laboran 432 empleados.

41. La dirección de ventas desea conocer, con un nivel de confianza de 90% y un margen de error de 8%, la proporción de asistentes a un concierto que adquirieron su boleto vía internet; por estudios anteriores, se considera que 2 de cada 10 adquirieron boletos por internet. Calcule el tamaño mínimo requerido de la muestra, si al concierto asisten 14 000 personas.

8.8 Otros intervalos de confianza

8.8.1 Estimación de diferencia de medias

42. Una fábrica lleva a cabo un análisis para definir a qué proveedor comprará su materia prima durante los siguientes 2 años y por ello desea saber la diferencia del promedio de tiempo de entrega de la mercancía una vez ordenada entre los 2 proveedores, para esto, toma una muestra aleatoria de 60 ventas de cada proveedor y encuentra que, en el primero, el tiempo de entrega promedio es de 14 días con una desviación estándar de 2.8, y en cuanto al segundo proveedor, este tiempo en promedio es de 18 días con una desviación estándar de 2.2.

Estime con un nivel de confianza de 96% la diferencia del promedio de tiempo de entrega de mercancías entre los 2 proveedores.

43. Una empresa de productos de belleza obtiene la mayor parte de sus utilidades por las ventas en campo, el gerente de los vendedores en campo divide a sus empleados según la zona en la que venden y desea estimar la diferencia del promedio de ventas diarias que realiza cada vendedor entre las 2 zonas de mayores ventas, por lo que toma una muestra aleatoria de 32 empleados de la zona A y encuentra que en promedio realizan 44 ventas diarias con una desviación estándar de 3.7, y toma otra muestra aleatoria de 40 empleados correspondiente a la zona B y encuentra que el promedio de ventas fue de 56 diarias por persona con una desviación estándar de 4.2. Encuentre la diferencia del promedio de ventas diarias por persona entre los empleados de las zonas A y B de dicha empresa con un nivel de confianza de 96 por ciento.

44. Una productora de materiales para la construcción tiene 2 áreas de diseño muy importantes y cada una de ellas maneja diferentes materiales, el gerente de diseño desea saber la diferencia del promedio de presupuesto que gastan estas 2 áreas diariamente, por lo que toma una muestra aleatoria de 20 días en cada una de las áreas y encuentra que en la primera se gasta en promedio $1 750 diarios con una des-

viación estándar de $350 y en la otra área el promedio era de $2 300 diarios con una desviación estándar de $375. Con un nivel de confianza de 92%, estime la diferencia de promedio del dinero utilizado diariamente entre las 2 áreas de diseño de esta empresa.

45. Una empresa de telefonía celular ofrece 2 tipos de paquetes a sus clientes: el básico y el premier; la gerencia de atención al cliente desea conocer la diferencia del promedio de quejas que se reciben semanalmente por parte de los usuarios de cada paquete, así que toma una muestra aleatoria de 35 semanas para cada paquete, y encuentra que, en el caso del paquete básico, se presentan 24 quejas con una desviación estándar de 3, y del paquete premier se presentan 13 quejas con una desviación estándar de 2. Con un nivel de confianza de 95%, estime la diferencia de promedio de quejas que se presentan para cada uno de los paquetes que ofrece la empresa.

46. Una hipotecaria desea conocer la diferencia del promedio de deuda que tienen los clientes de 2 diferentes zonas del país, para ello, toma una muestra aleatoria de la zona A de 25 personas y encuentra que en promedio la deuda asciende a $728 300 con una desviación estándar de $27 850, y toma una muestra aleatoria de 35 para la zona B y encuentra que la deuda promedio es de $597 450 con una desviación estándar de $37 500. Con un nivel de confianza de 98%, estime la diferencia de promedio la deuda que tienen las personas de las 2 zonas.

8.8.2 Estimación de diferencia de proporciones

47. Una fábrica de neumáticos tiene 2 máquinas que realizan el mismo proceso y se desea saber la diferencia entre las proporciones de artículos defectuosos que genera cada una de ellas, por lo que se toma una muestra aleatoria de 300 neumáticos de cada máquina y encuentra que en la primera hay 12 y en la segunda 18 artículos defectuosos. Estime, con un nivel de confianza de 95%, la diferencia entre las proporciones de neumáticos defectuosos que genera cada máquina.

48. El gobierno realiza un estudio sobre las causas de fracaso de microempresas en los 2 estados con mayor actividad económica del país, obtuvo una muestra aleatoria de 250 microempresas de cada estado y encontró que 170 del primer estado y 140 del segundo, tenían activos fijos superiores a 65% del valor total de la empresa en el momento del fracaso. Con un nivel de confianza de 98%, estime la diferencia entre las proporciones de microempresas que fracasaron entre los 2 estados.

49. Dos empresarios tienen un proyecto para ofrecer servicios de limpieza profesional para que funcione como *outsourcing* de otras empresas. Monterrey y Guadalajara son los primeros mercados en los que planean incursionar; uno de ellos cree que es irrelevante la ciudad por lo que el segundo quiere saber la diferencia de proporciones de empresas que contratan la limpieza por *outsourcing* de cada ciudad; con una muestra aleatoria de 55 empresas de cada lugar, se obtiene que en Monterrey 38 de ellas contratan aparte el servicio y, en Guadalajara, 27. Con un nivel de confianza de 90%, estime la diferencia entre estas proporciones y determinar si la ciudad para llevar a cabo el proyecto realmente es irrelevante.

50. El nuevo gerente del área de finanzas de una empresa comercial desea tener un mejor control sobre los pagos a proveedores y cobros a clientes de acuerdo con las políticas de la empresa, por lo que toma una muestra aleatoria de 40 clientes y otra de 40 proveedores y observa que 85% de los clientes cumplen con sus pagos a tiempo, mientras que la empresa sólo paga a 73% de sus proveedores en el plazo establecido. Estime, con un nivel de confianza de 95%, la diferencia entre estas proporciones teniendo en cuenta que debería ser mínima.

51. El área de crédito y cobranzas de un banco desea saber la diferencia entre las proporciones de sus clientes de 18 a 40 años y aquellos mayores de 40 en relación con cuántos de éstos exceden el límite de crédito que se les otorga, para investigarlo, toma una muestra aleatoria de 220 de cada grupo de clientes y encuentra que, de aquellos entre 18 y 40 años, 121 sobrepasaban su límite, mientras que de los clientes mayores de 40 años, a sólo 66 les ocurría esto. Con un nivel de confianza de 96%, estime la diferencia de proporciones que existe entre un grupo y otro.

8.8.3 De un total a partir de una media

52. Una cadena de pizzas a domicilio que cuenta con 120 tiendas a lo largo del país tomó una muestra de 50 tiendas y se registró el número de pedidos que recibieron durante un día, después se estimó, con nivel de confianza de 80%, que el promedio de pedidos ese día estuvo entre 70.86 y 77.14. Construya un intervalo de confianza para el total de pedidos en todas las tiendas.

53. De 2 000 alumnos que asisten a una escuela secundaria se tomó una muestra de 200 a los que se les preguntó el número de horas que dedican a la semana para practicar algún deporte, se estimó, con un nivel de confianza de 85%, que el número promedio de horas que dedican al deporte está entre 10.04 y 28.06. Construya un intervalo de confianza para el total de estudiantes.

54. De los 300 días al año que trabaja una imprenta, se tomó el número de impresiones a blanco y negro que se realizaron durante una muestra aleatoria de 90 días, se estimó, con un nivel de confianza de 95%, que el promedio diario de impresiones a blanco y negro está entre 26.11 y 38.15. Construya un intervalo de confianza para el total de días que trabaja.

8.8.4 De un total a partir de una proporción

55. Una compañía que elabora helados desea estimar, con una confianza de 95%, la proporción de niños de entre 8 y 10 años que prefieren el sabor a chocolate; se tomó una muestra de 150 niños entre esas edades y se encontró que la proporción de niños que prefieren el helado de ese sabor está entre 50 y 66%. Construya un intervalo de confianza para el total de niños que prefieren el helado de chocolate, sabiendo que el mercado de esa empresa consta de 565 000 niños de esas edades.

56. Se desea estimar, con una confianza de 99%, la proporción de la población de consumidores de cierta marca de champú que adquiere la presentación de 500 ml; se tomó una muestra aleatoria de 100 de esos consumidores y se encontró que la proporción de consumidores que adquieren esa presentación está entre 25 y 49%. Construya un intervalo de confianza para el total de consumidores del champú que compran la presentación de 500 ml sabiendo que existen 234 600 personas que consumen la marca.

57. El departamento de limpia de la ciudad de México desea saber la proporción de hogares que separan la basura con un nivel de confianza de 99%; se realizaron 500 entrevistas al azar y se estimó que la proporción de hogares que sí lo hacen se encuentra ente 14 y 22%. Construya un intervalo de confianza para el total de hogares sabiendo que existen 2 180 240 de hogares en la ciudad de México.

58. Se desea conocer, con un nivel de confianza de 85%, la proporción de la población de una ciudad grande que viaja en metro, se encuestaron a 600 personas al azar y se encontró que la proporción que usa el metro está entre 40 y 46%. Construya un intervalo de confianza para el total de personas que utilizan el metro sabiendo que en la ciudad existen 3 250 320 habitantes.

59. Se desea conocer, con un nivel de confianza de 94%, la proporción de los alumnos que ingresan a licenciatura en una institución universitaria contando con conocimientos de un idioma adicional; se tomó una muestra aleatoria de 200 alumnos y se encontró que la proporción que tiene conocimientos de otro idioma está entre 16 y 26%. Construya un intervalo de confianza para el total de estudiantes sabiendo que ingresaron 1 732.

capítulo 9

Pruebas de hipótesis

Sumario

9.1 Introducción
9.2 Planteamiento de las hipótesis
9.3 Errores tipo I y tipo II
9.4 Procedimiento para realizar pruebas de hipótesis
9.5 Elaboración de una gráfica
9.6 Pruebas de 1 y de 2 extremos. Regiones de aceptación y de rechazo
9.6.1 Pruebas de 2 extremos o colas
9.6.2 Prueba de hipótesis de la cola inferior o del extremo izquierdo
9.6.3 Prueba de hipótesis de la cola superior o del extremo derecho
9.7 Métodos para realizar pruebas de hipótesis
9.7.1 Método del intervalo
9.7.2 Método del estadístico de prueba
9.7.3 Método del valor de la *P*
9.7.4 Resumen de los procedimientos para realizar pruebas de hipótesis con los 3 métodos
9.8 Prueba de hipótesis sobre una proporción poblacional
9.9 Resumen
9.10 Excel: Uso de Excel
9.11 Ejercicios adicionales

9.1 Introducción

Una **hipótesis estadística** es una suposición o afirmación sobre alguna característica de una población. Por ejemplo, "el promedio de contenido de las cajas del cereal X es de 300 g" es una afirmación sobre el promedio de peso, en gramos, de esas cajas.

Hipótesis estadística. Es una suposición o afirmación sobre alguna característica de una población.

El interés en una hipótesis como la anterior puede abordarse desde diferentes perspectivas. El fabricante podría probar si las cajas de cereal realmente tienen ese contenido promedio, o algún valor cercano, para determinar si su proceso de producción está funcionando adecuadamente. Esto permitiría saber si el gerente de producción hace bien su trabajo. A un competidor le interesaría saber si ese valor es correcto para orientar su estrategia de competencia; mientras que a una agencia de protección al consumidor le interesaría comprobar si el fabricante cumple con lo que ofrece a sus clientes. Por supuesto, la hipótesis se plantea desde el punto de vista de quien está evaluando la situación.

El ejemplo del contenido de las cajas de cereal es una hipótesis sobre un parámetro poblacional, es decir, el valor de la media de una población. El planteamiento formal de esa hipótesis es el siguiente: H_0: $\mu = 300$ en donde H_0 es lo que se denomina **hipótesis nula**, es decir, la hipótesis que desea probarse. También es común que se planteen hipótesis sobre otros parámetros poblacionales como: la mediana de una población o π, la proporción de una población, o σ^2, la varianza de una población. Asimismo se formulan hipótesis sobre la diferencia de las medias, las proporciones o las varianzas de 2 poblaciones, $\mu_1 - \mu_2$, $\pi_1 - \pi_2$ y $\sigma_1^2 - \sigma_2^2$, respectivamente. De la misma manera se prueban hipótesis sobre la posible igualdad entre más de 2 medias o varianzas. A este tipo de pruebas se les conoce como pruebas de hipótesis paramétricas, precisamente porque se realizan sobre parámetros poblacionales.

Hipótesis nula. Es la hipótesis que desea probarse.

Igualmente se plantean hipótesis sobre características que no son parámetros poblacionales. Se prueban hipótesis sobre **bondad de ajuste**, es decir, sobre qué tan bien se ajusta una serie de datos a una distribución normal, binomial u otro tipo de distribución. Existen también hipótesis sobre la independencia entre 2 o más variables o sobre la aleatoriedad de un conjunto de datos observados como, por ejemplo, una serie de resultados de algún proceso aleatorio como un juego de azar. Este tipo de pruebas se llaman *pruebas de hipótesis no paramétricas* y se analizan con mayor detalle en los capítulos 11 y 17, en tanto que en este capítulo y el siguiente se estudian los principales tipos de pruebas de hipótesis paramétricas. Además, en los capítulos 12 y 13 se revisarán otros ejemplos de pruebas de hipótesis paramétricas que se utilizan en los temas de análisis de varianza y de análisis de regresión, objeto de estudio de estos dos capítulos.

Bondad de ajuste. Grado de ajuste que tiene una serie de datos a una distribución normal, binomial o de otro tipo.

9.2 Planteamiento de las hipótesis

El procedimiento de pruebas de hipótesis consiste básicamente en que una vez planteada la hipótesis que desea probarse, se prosigue a recabar datos muestrales para conocer si éstos concuerdan con ella o si la contradicen. En el primer caso, si no hay información que permita refutar la hipótesis, se considera que es válida o, más estrictamente, que no hay razón para rechazarla. Si por el contrario la información muestral no concuerda con lo que la hipótesis plantea, entonces se rechaza para concluir que lo contrario es cierto.

En el ejemplo de las cajas de cereal, la hipótesis que desea probarse es que el contenido promedio de las cajas es de 300 g, de tal manera que se formula lo que se conoce como **hipótesis nula**: H_0: $\mu = 300$. Si esta afirmación o hipótesis no es verdadera entonces resulta que lo contrario es lo cierto, es decir, que el contenido promedio de las cajas no es de 300 g. En símbolos: H_1: $\mu \neq 300$. Y H_1 es lo que se denomina **hipótesis alternativa**, y es la que se asume como verdadera en caso de que la nula resulte no serlo.

Hipótesis nula. Es la afirmación o hipótesis que no es verdadera.
Hipótesis alternativa. Es la hipótesis que se asume como verdadera en caso de que la nula resulte no serlo.

Es importante analizar con detenimiento cómo se plantean las hipótesis. Después de H_0 y H_1 están dos puntos (:) que indican que a continuación se presenta la hipótesis. Esto es importante porque suele cometerse el error de sustituir los dos puntos por el signo igual que (=), y aunque pudiera ser válido se presta a confusiones con los signos de igualdad que se utilizan en la hipótesis misma ($\mu = 300$, la nula, o $\mu \neq 300$, la alternativa).

9.3 Errores tipo I y tipo II

Debido a que la decisión de aceptar o rechazar la hipótesis nula se hace con base en datos muestrales, no se tiene una confiabilidad de 100% de que la decisión sea correcta por lo que, al igual que se hizo con la estimación de parámetros en el capítulo anterior, a la decisión se le asigna un determinado nivel de confianza.

Para explicar cómo se maneja ese nivel de confianza es necesario revisar los 4 posibles resultados de la decisión: se puede aceptar o rechazar la hipótesis nula cuando ésta puede ser cierta o falsa, esto a su vez conduce a 2 tipos de decisiones correctas y a 2 incorrectas. En el cuadro siguiente se resumen estas consideraciones:

	H_0 es verdadera	H_0 es falsa
Se acepta H_0	Correcto	Error tipo II
Se rechaza H_0	Error tipo I	Correcto

Error tipo I. Error que consiste en rechazar una hipótesis que es verdadera.
Error tipo II. Error que consiste en aceptar una hipótesis falsa.

En este cuadro se observa claramente cuáles son las 2 decisiones correctas: aceptar una H_0 que es cierta o rechazar una que es falsa. Por otra parte, al error que consiste en rechazar una hipótesis que es verdadera se le denomina **error tipo I** y al que consiste en aceptar una hipótesis falsa se le llama **error tipo II**.

Además, se utiliza la letra griega alfa (α) para representar la probabilidad de cometer el error tipo I cuando se realiza una prueba de hipótesis y la beta (β) para representar la probabilidad de cometer el error tipo II.

En el ejemplo anterior de las cajas de cereal se tenían las 2 siguientes hipótesis:

$$H_0: \mu = 300$$
$$H_1: \mu \neq 300$$

Supóngase ahora que por estudios anteriores se sabe que la desviación estándar del proceso de producción completo de esas cajas de cereal (la población) es de 15 g. Si se utiliza esta información para controlar el proceso de producción podrían tomarse muestras, por ejemplo de 36 cajas de cereal, y evaluar si el proceso está o no bajo control; es decir, saber si el promedio del contenido de las cajas de cereal es o no de 300 g, con lo que se tomarían las decisiones pertinentes: dejar que el proceso continúe en caso de que la producción esté saliendo bien o, de lo contrario, detenerlo y tomar las medidas correctivas necesarias. Supóngase que al tomar una muestra se encuentra que su promedio de contenido es menor de 295 g o mayor de 305 se asume entonces que no se están produciendo cajas de cereal con contenido promedio de 300 g; por el contrario, si el promedio de la muestra se encuentra entre esos 2 límites, se asumirá que el proceso

está trabajando bien. Con esta información y presumiendo que la hipótesis nula es cierta, se tendrían las condiciones que se ilustran en la figura 9.1

Como puede verse en la figura, si el verdadero promedio de contenido de las cajas de cereal es de 300 g con una desviación estándar de 15 g, la desviación estándar de la distribución muestral (el error muestral) para muestras de tamaño $n = 36$ es:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{36}} = 2.5$$

Con este valor calculamos, a partir de la tabla de áreas bajo la curva normal, que la proporción de todas las muestras posibles de tamaño 36 (la distribución muestral) que tienen peso de entre 295 y 305 g es la siguiente:

$$z = \frac{X - \mu}{\sigma_{\bar{X}}} = \frac{295 - 300}{2.5} = -2$$

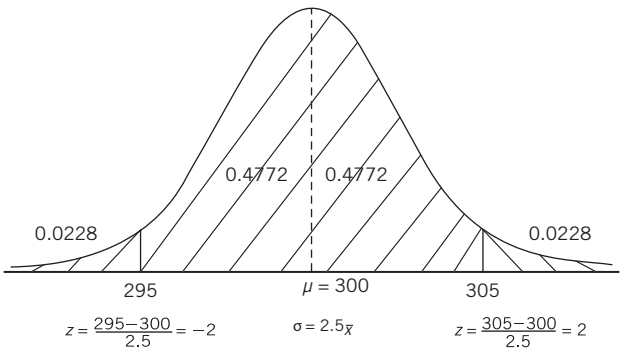


Figura 9.1 Condiciones para el ejemplo del promedio de contenido de cajas de cereal.

De la tabla de áreas bajo la curva normal se tiene que:

$$P(-2 \le z \le 0) = P(295 \le X \le 300) = 0.4772$$

Dada la simetría de la curva normal, esta probabilidad es la misma para la proporción de muestras cuyo promedio está entre 300 y 305 g:

$$P(0 \le z \le 2) = P(300 \le X \le 305) = 0.4772$$

De lo anterior se resuelve que la probabilidad que tiene una muestra aleatoria de 36 cajas de cereal con un promedio de peso entre 295 y 305 g es de:

$$(0.4772)(2) = 0.9544$$

En contraparte, la probabilidad de que una muestra elegida al azar pese menos de 295 g es:

$$P(-\infty \le z \le -2) = P(X \le 295) = 0.5 - 0.4772 = 0.0228$$

Y la probabilidad de que pese más de 305 g es:

$$P(2 \le z \le \infty) = P(X \ge 305) = 0.5 - 0.4772 = 0.0228$$

Por lo que la probabilidad de que una muestra, elegida al azar de esa distribución muestral, pese menos de 295 o más de 305 g es:

$$P(X < 295) + P(X > 305) = 0.0228 + 0.0228 = 0.0456, \text{ o } 4.56\%$$

La cual es una probabilidad relativamente reducida. De esto se desprende que, al calificar de inadecuado el proceso de empacado cuando el promedio del peso de las cajas de cereal es menor de 295 o mayor de 305 g, se está aceptando una probabilidad de 4.56% de estar cometiendo el error tipo I, lo que significa rechazar una hipótesis verdadera porque existe la probabilidad de que el peso promedio de una muestra aleatoria de 36 cajas exceda esos límites, cuando el verdadero promedio es de 300 g. A ese 4.56%, que corresponde a la probabilidad aceptada de estar cometiendo el error tipo I, se le denomina α, y se le conoce como *nivel de significación*. Se resumen todas estas consideraciones en la figura 9.2.

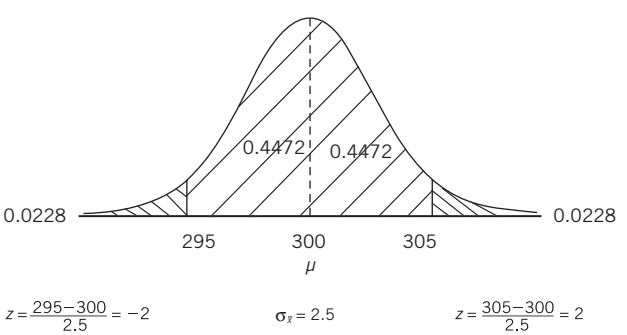


Figura 9.2 α, el nivel de significación.

Para ilustrar ahora β, la probabilidad de estar cometiendo el error tipo II mismo que consiste en aceptar una hipótesis falsa, supóngase que con la misma regla de decisión anterior —aceptar H_0 si el valor observado está entre 295 y 305 g— ocurre que el verdadero promedio de peso del cereal de esas cajas no es de 300 g sino de 290 g, por lo tanto se tiene que:

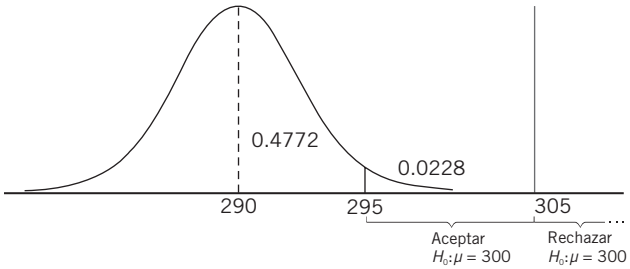


Figura 9.3 Peso promedio de cajas de cereal de 290 g.

$$z = \frac{\bar{X} - \mu}{\sigma_{\bar{x}}} = \frac{295 - 290}{2.5} = 2$$

Y,

$$P(z \geq 2) = 0.0228$$

Se ilustran las condiciones en la figura 9.3.

Como puede verse en esta figura, si el peso promedio de las cajas es de 290 g, la probabilidad de que el peso promedio de una muestra aleatoria de 36 de ellas sea mayor de 295 g es de 2.28%, y si la muestra que se extrae pesa más de esta cantidad, de acuerdo con la regla de decisión adoptada (aceptar H_0 si el valor observado está entre 295 y 305 g), se aceptaría la hipótesis nula de que $\mu = 300$, lo cual constituye un error tipo II ya que se aceptaría una hipótesis falsa. En este caso se asumiría una β, probabilidad de cometer el error de aceptar una hipótesis falsa, de 2.28%. Ahora bien, ¿qué sucede si esta μ verdadera no es de 290 sino que es de 292.5? En este caso:

$$z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \frac{295 - 292.5}{2.5} = 1$$

Y,

$$P(z \geq 1) = 0.1587.$$

Se ilustran las condiciones en la figura 9.4.

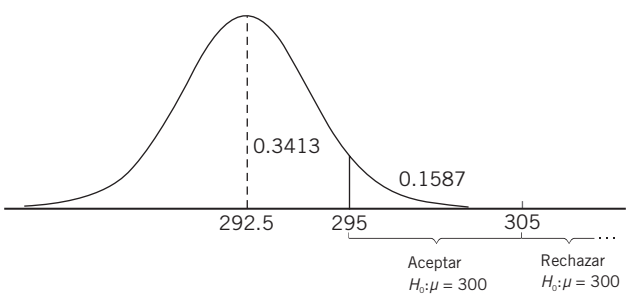


Figura 9.4 Peso promedio de cajas de cereal de 292.5 g.

En este caso si μ, el peso promedio de las cajas, es de 292.5 g la probabilidad de que el peso promedio de una muestra aleatoria de 36 de ellas sea mayor de 295 g es de 15.87%; si la muestra que se extrae pesa más que esta cantidad, de acuerdo con la regla de decisión adoptada (aceptar H_0 si el valor observado está entre 295 y 305 g), se aceptaría la hipótesis nula de que $\mu = 300$, y se cometería el error de aceptar una hipótesis falsa.

Como puede verse el valor de β, la probabilidad de cometer el error de aceptar una hipótesis falsa, depende de que el valor del parámetro planteado en la hipótesis nula ($\mu = 300$ en este caso) no sea el verdadero y también de qué tan cercano o distante esté ese valor verdadero.

Se ilustraron 2 posibles valores alternativos para la verdadera media del contenido de cereal de las cajas, 290 y 292.5, pero cabe aclarar que es infinita la cantidad de valores que podría asumir la verdadera media. Para evaluar cómo se comporta esta β, en la tabla 9.1 se resumen los valores correspondientes a β para diversos valores posibles de μ.

Tabla 9.1 Valores de β para diversos valores posibles de μ

μ	z	P
290.00	2	0.0228
290.50	1.8	0.0359
291.00	1.6	0.0548
291.50	1.4	0.0808
292.00	1.2	0.1151
292.50	1	0.1587
293.00	0.8	0.2119
293.50	0.6	0.2743
294.00	0.4	0.3446
294.50	0.2	0.4207
295.00	0	0.5000

El procedimiento para determinar estas probabilidades es el mismo que se ilustró con anterioridad. Nótese que las probabilidades para $\mu = 290$ y $\mu = 292.50$ son las mismas que se encontraron en los 2 ejemplos previos.

En la figura 9.5 se ilustra una curva suave con estos valores la cual se conoce como **curva característica operativa** o CCO; ésta muestra los niveles de riesgo, las probabilidades de cometer el error de aceptar una hipótesis falsa, para diversos valores hipotéticos de la verdadera media poblacional.

Curva característica operativa. Muestra los niveles de riesgo, las probabilidades de error al aceptar una hipótesis falsa para diversos valores hipotéticos de la verdadera media poblacional.

Esta curva característica operativa presenta las cualidades de un posible criterio de prueba y, entre otros detalles, muestra que la probabilidad de aceptar la hipótesis nula es más alta precisamente cuando es cierta; además explica que esa probabilidad también es alta cuando el verdadero valor de la población está ligeramente alejado del valor planteado por la hipótesis nula.

Asimismo la curva hace notar que para valores de la media cada vez más alejados del valor supuesto por la hipótesis nula la probabilidad de detectar esas diferencias y de aceptar equivocadamente la hipótesis es cada vez menor, lo cual es deseable.

Un estudio detallado de las curvas características operativas rebasaría el alcance de este libro cuyo principal propósito es mostrar cómo pueden aplicarse métodos estadísticos para medir y controlar riesgos en los que se incurre al realizar pruebas de hipótesis.

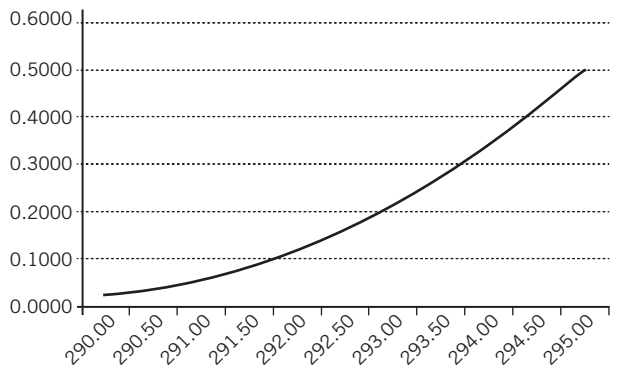


Figura 9.5 Curva característica operativa.

Para cerrar el tema de los errores tipo II vale la pena repasar qué sucede con las auditorías contables. Es sabido que una auditoría debe concluir si los estados financieros de las empresas reflejan razonablemente su información financiera. En muchos casos esta decisión se toma con base en la información muestral. En este tipo de situaciones resulta más grave el error tipo II (aceptar que la información financiera es adecuada cuando en realidad no lo es) que el error tipo I (concluir que la información no es adecuada cuando en verdad sí lo es). Por ello en auditoría es especialmente importante tomar en cuenta ese error tipo II.

9.4 Procedimiento para realizar pruebas de hipótesis

Se ha revisado el tema de las hipótesis nula y alternativa así como el de los tipos de errores que pueden cometerse al realizar pruebas de hipótesis. En seguida se plantea un ejemplo común de prueba de hipótesis para utilizarlo como base al ilustrar este procedimiento, tanto como para explicar otros conceptos importantes relacionados con pruebas de hipótesis estadísticas.

EJEMPLO 9.1

Un fabricante de llantas para automóvil afirma que la duración promedio de determinado modelo de llanta en un coche de cierto peso es de 40 000 km bajo condiciones normales de manejo. Se analiza una muestra aleatoria de 100 llantas de este tipo bajo las condiciones especificadas, y se encuentra que la duración promedio fue de 39 000 km con una desviación estándar de 8 500 km. Con esta información y con lo que se ha revisado sobre distribuciones muestrales, incluidos los planteamientos básicos del teorema central del límite, realice una prueba de hipótesis estadística para decidir si la afirmación del fabricante es aceptable o no.

El primer paso a seguir es plantear formalmente la hipótesis a probar. En símbolos:

$$H_0: \mu = 40\,000$$

Así se representa lo que el fabricante afirma. Tal como se mencionó se requiere también una hipótesis alternativa en caso de que los datos muestrales no concuerden con aquella afirmación. La hipótesis alternativa es:

$$H_1: \mu \neq 40\,000$$

Ésta resume en símbolos la conclusión a la que se llegaría si se decide que la afirmación del fabricante no es sostenible.

En seguida se establece un nivel de significación alfa (α) que representa la máxima probabilidad aceptada de cometer el error de tipo I, es decir, rechazar la hipótesis si es verdadera. Los valores del nivel de significación más comunes son 0.01 y 0.05, aunque también podría ser cualquier otro valor siempre y cuando sea reducido, ya que es deseable una probabilidad baja de cometer el error de rechazar una hipótesis verdadera. En este

ejemplo se utilizará $\alpha = 0.05$. Si la hipótesis nula es verdadera y la muestra de 100 llantas es aleatoria (representativa de la población) entonces se sabe que la media muestral observada ($\overline{X} =$ 39 000) pertenece a una distribución muestral de tipo normal (en forma de campana); ésta tiene como media a la verdadera media poblacional y como desviación estándar el correspondiente error estándar de la media:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N - n}{N - 1}}$$

En este caso, como se trata de una población muy grande (el gran número de llantas que se fabrican) se reduce a:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Si se utiliza la desviación estándar de la muestra como estimador de la desviación estándar de la población, entonces se puede calcular este error estándar a partir del estadístico muestral:

$$s_{\bar{X}} = \frac{s}{\sqrt{n}} = \frac{8\,500}{\sqrt{100}} = \frac{8\,500}{10} = 850$$

Con este error estándar puede ilustrarse gráficamente la situación que plantea la hipótesis nula, tal como la figura 9.6 ejemplifica.

Como $P(-1.96 \leq z \leq 1.96) = 0.05$, 95% de todas las medias posibles se encuentran dentro del intervalo 38 334 a 41 666 km

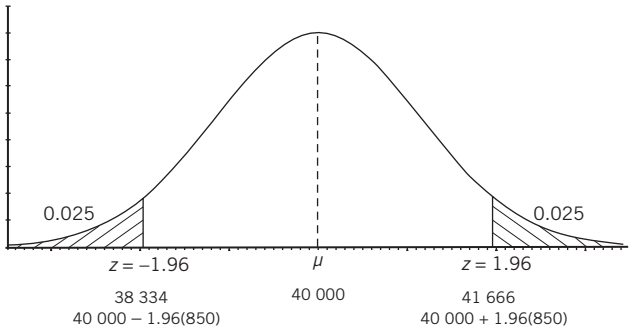


Figura 9.6 Datos del ejemplo 9.1 en una gráfica de distribución normal.

y la regla de decisión indica rechazar la hipótesis nula sólo si el valor observado de la media muestral está fuera de este intervalo, en este caso no se rechaza H_0, y se concluye que el promedio de duración de las llantas observado en la muestra es consistente con la afirmación del fabricante, lo cual quiere decir que sí puede afirmarse que la duración promedio de ese modelo de llanta es de 40 000 km bajo condiciones normales de manejo en un automóvil del peso especificado.

Aunque ese intervalo contiene 95% de todos los valores posibles de las medias muestrales cualquier otro valor es posible, sin embargo es tan reducida su probabilidad de ocurrencia que si se obtuviera un valor fuera del intervalo la hipótesis nula sería rechazada y se incurriría en una probabilidad de 5% de rechazar una hipótesis verdadera, el nivel de significación especificado.

En resumen, en el procedimiento anterior los pasos a seguir para realizar pruebas de hipótesis son los siguientes:

1. Plantear las hipótesis nula y alternativa.
2. Establecer el nivel de significación al que se desea realizar la prueba.
3. Calcular el error estándar del estadístico (en el ejemplo el error estándar de la media).
4. Con base en el nivel de significación y en el error estándar del estadístico, establecer la regla de decisión, es decir, los valores dentro de los que se acepta la hipótesis nula y aquellos en los que se rechaza. Como se verá más adelante, estos valores que permiten aceptar o rechazar la hipótesis nula (y en contraparte aceptar o rechazar la alternativa) dependen de cuál de los 3 métodos de prueba se aplica: *a*) intervalo, *b*) estadístico de prueba o *c*) probabilidad.
5. Tomar la decisión con base en todos los elementos anteriores.
6. Finalmente, de suma importancia, interpretar los resultados en términos del planteamiento original.

En las secciones siguientes se revisarán con mayor detalle otros aspectos importantes de las pruebas de hipótesis al tiempo que seguirá aplicándose este procedimiento.

9.5 Elaboración de una gráfica

Se decidió insertar esta brevísima sección para incorporar una recomendación que suele ser sumamente útil: ilustrar en una gráfica los datos involucrados en una prueba de hipótesis. Estas gráficas son una ayuda muy valiosa, simple pero muy útil, para visualizar las circunstancias de una prueba de hipótesis.

EJERCICIOS 9.5 Introducción

1. ¿Qué es una hipótesis?
2. ¿Qué es una hipótesis estadística?
3. ¿Cuáles son los 2 tipos de hipótesis utilizadas en las pruebas de hipótesis estadísticas?
4. Proporcione algunos ejemplos de parámetros sobre los que suelen plantearse hipótesis estadísticas.
5. Además de las pruebas de hipótesis sobre parámetros ¿qué otros tipos de hipótesis estadísticas suelen utilizarse?
6. Explique el mecanismo básico que permite aceptar o rechazar una hipótesis estadística.
7. ¿Cuáles son los tipos de decisiones correctas que pueden tomarse cuando se realizan pruebas de hipótesis?
8. ¿Cuáles son los tipos de decisiones equivocadas que pueden tomarse cuando se realizan pruebas de hipótesis?
9. ¿En qué consiste el error tipo I en una prueba de hipótesis?
10. ¿En qué consiste el error tipo II en una prueba de hipótesis?
11. En una prueba de hipótesis ¿qué es α?
12. En una prueba de hipótesis ¿qué es β?
13. ¿Qué es una curva característica operativa?
14. Explique el procedimiento para realizar pruebas de hipótesis.
15. ¿Por qué es conveniente utilizar gráficas cuando se realizan pruebas de hipótesis?

9.6 Pruebas de 1 y de 2 extremos. Regiones de aceptación y de rechazo

En esta sección se revisan estos importantes conceptos en subsecciones separadas. Los 3 casos que pueden presentarse en pruebas de hipótesis y que se desprenden de la forma en éstas se plantean: pruebas de 2 extremos y de un extremo, del extremo derecho y del izquierdo.

9.6.1 Pruebas de 2 extremos o colas

En la figura 9.6 puede verse que el área total de la distribución normal queda dividida visiblemente en 2 áreas pero son 3:

1. a la izquierda de $\overline{X} = 38\,334$
2. entre este valor y $\overline{X} = 41\,666$
3. el área a la derecha de $z = 1.96$

Sin embargo, puede considerarse que son 2 áreas para efectos de la prueba de la hipótesis. Las 2 áreas que se encuentran en ambos extremos de la distribución puede denominarse **región de rechazo**, ya que si el valor de la media muestral estuviera en alguna de ellas se refutaría la hipótesis nula. En otras palabras y símbolos se rechazaría H_0, si $\overline{X} < 38\,334$ o si $\overline{X} > 41\,666$. Por un razonamiento análogo, el área entre estos 2 valores es la **región de aceptación** porque, como sucedió en el ejemplo, si el valor observado de la media muestral 39 000 está entre 38 334 y 41 666 no se rechaza la hipótesis nula y se concluye que el fabricante tiene razón al afirmar que el promedio de vida de sus llantas es de 40 000 km.

Región de rechazo. Son las 2 áreas que se encuentran en ambos extremos de la distribución.
Región de aceptación. Es el área que se encuentra alrededor de la media y limitada por los valores críticos de los extremos de la distribución.

Las regiones de aceptación y de rechazo para el ejemplo 9.1 se ilustran en la figura 9.6; aparecen de esa manera debido a la forma de las hipótesis nula y alternativa que implicaban una igualdad y una diferencia ($=$ y $\neq$), respectivamente. En los 2 ejemplos siguientes se ilustran pruebas de hipótesis que implican desigualdades ($\leq$ y $\geq$) que ocasionan cambios en la forma cómo se definen las áreas de aceptación y de rechazo.

9.6.2 Prueba de hipótesis de la cola inferior o del extremo izquierdo

En seguida se revisan algunos ejemplos de la prueba de hipótesis de la cola inferior o del extremo izquierdo.

EJEMPLO 9.2

Fumar cigarros de la marca X produce en promedio 0.6 mg de nicotina. El departamento de ingeniería del fabricante propone un filtro nuevo que supuestamente reduce la producción de nicotina. Se toma una muestra de 50 cigarros con el nuevo filtro y se encuentra que el promedio de nicotina es de 0.55 mg con desviación estándar de 0.56. ¿Se debe aceptar la aseveración del

departamento de ingeniería con un nivel de significación de 2.5 por ciento?

Solución: En primer lugar, las hipótesis:

$$H_0: \mu = 0.60$$
$$H_1: \mu < 0.60$$

En la figura 9.7 se ilustran las circunstancias.

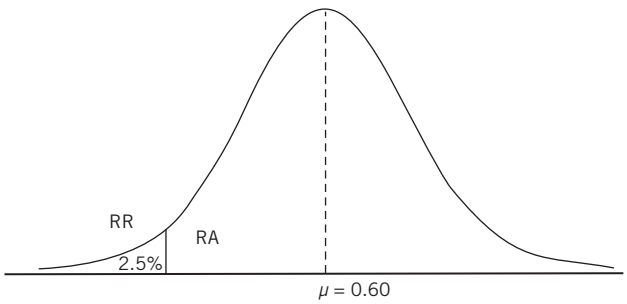


Figura 9.7 Circunstancias para el ejemplo 9.2.

Se trata de probar que los cigarros producen menos nicotina, entonces, la región de rechazo se ubica del lado izquierdo de la campana, en donde se divide 2.5% del nivel de significación deseado y se rechaza H_0 si el valor encontrado en la muestra es lo suficientemente bajo como para contradecirla y apoyar, en cambio, lo planteado por la hipótesis alternativa. El valor de z que aísla 2.5% del área a la izquierda es –1.96; en símbolos:

$$P(-\infty \leq z \leq -1.96) = 0.025, \text{ en términos de } z.$$

Para expresar esta misma probabilidad en términos de la variable, la nicotina, se calcula el error estándar:

$$s_{\bar{X}} = \frac{s}{\sqrt{n}} = \frac{0.56}{\sqrt{50}} = \frac{0.56}{7.071}$$
$$= 0.0792$$

Y,

$$P(X < 0.60 - 0.0792(1.96)) = P(X < 0.60 - 0.155)$$
$$= P(X < 0.0.445) = 0.025$$

Por lo anterior, la regla de decisión consiste en aceptar H_0 si el valor observado en la muestra es igual o mayor que 0.445 y en rechazarla si es menor. Como el valor observado de la muestra es de un promedio de 0.55 mg de nicotina, cae en la región de aceptación y, por lo tanto, no se rechaza la hipótesis nula, así se concluye que el nuevo filtro no reduce la generación de nicotina en esos cigarros, los cuales siguen teniendo un promedio de 0.60 mg.

9.6.3 Prueba de hipótesis de la cola superior o del extremo derecho

Ejemplo 9.3

Una institución de enseñanza de idiomas registra en promedio 4 personas semanalmente en los cursos que ofrece, por lo que hace una campaña publicitaria de gran magnitud. Después de un año, toma una muestra aleatoria de 32 semanas y encuentra que en promedio se registran semanalmente 5 alumnos nuevos con una desviación estándar de 0.9. Tomando en cuenta un nivel de significación de 0.05, determine si la campaña publicitaria cumplió su objetivo.

Solución: Las hipótesis:

$$H_0: \mu = 4$$
$$H_1: \mu > 4$$

Estas hipótesis se plantean así porque si el aumento en el registro de alumnos nuevos no es significativo, entonces se asume que no hubo incremento y que lo observado en la muestra se debe simplemente al azar.

El error estándar de la media:

$$s_x = \frac{s}{\sqrt{n}} = \frac{0.9}{\sqrt{32}} = \frac{0.9}{5.6568} = 0.1591$$

El valor de z que aísla 5% del área bajo la curva es 1.645, por lo que en términos del número de alumnos que se inscribe:

$$\mu + s_x(z)$$
$$4 + 0.1591\ (1.645)$$
$$4 + 0.2617 = 4.2617$$

En la figura 9.8 se representan estos datos.

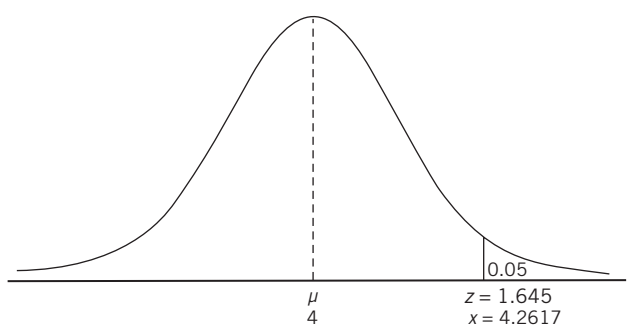


Figura 9.8 Datos del ejemplo 9.3.

Como puede observarse, un valor de 4.2617 en el número de alumnos inscritos divide la región de rechazo a la izquierda y la de aceptación a la derecha. De acuerdo con los datos de la muestra, la inscripción semanal es de 5 personas y como ese número es mayor a 4.2617 la hipótesis nula es rechazada y se concluye que la campaña publicitaria cumplió su objetivo.

Para resolver los ejemplos anteriores se utilizó el denominado *método del intervalo*. En la sección siguiente se revisarán 2 métodos más que pueden utilizarse para realizar pruebas de hipótesis.

EJERCICIOS 9.6 Pruebas de 1 y de 2 extremos. Regiones de aceptación y de rechazo

1. Un estudio señaló que cierta marca de baterías tipo AA tiene una duración promedio de 825 horas de uso continuo. Al analizar una muestra de 120 baterías AA se encuentra que la duración promedio es de 805 horas con una desviación estándar de 12. Con base en los resultados observados en la muestra y con un nivel de significación de 0.05, ¿puede asegurarse que lo dicho en el estudio es correcto?
2. Una máquina expendedora de refrescos está programada para servir 250 ml por vaso. Una muestra de 50 vasos mostró que en promedio sirve 252 ml con una desviación estándar de 8 ml. ¿Puede asegurarse que la afirmación es correcta con un nivel de significación de 0.01?
3. Un fabricante asegura que el promedio de tiempo de duración del sabor de cierta goma de mascar es de 18 minutos. Una muestra de 85 piezas arrojó un promedio de duración del sabor de 15.7 minutos con una desviación estándar de 3.5 minutos. ¿Puede decirse que la afirmación del fabricante es correcta con un nivel de significación de 0.05?
4. Un estudio publicado recientemente afirma que en promedio una persona bebe 58 L de café al año. Al tomar una muestra de 110 personas se encontró que en promedio beben 60 L de café al año con una desviación estándar de 13. ¿Puede afirmarse que el estudio es correcto? Utilice un nivel de significación de 0.05.
5. El gerente general de un banco afirma que el promedio que se tarda un cliente en realizar sus operaciones bancarias es de 4.8 minutos. Se analiza una muestra de 45 operaciones bancarias y se descubre que en promedio se realizaron en 5.2 minutos, con una desviación estándar de 2.4 minutos. ¿Puede afirmarse que el gerente está en lo correcto? Utilice un nivel de significación del 0.01.

Prueba de hipótesis de la cola inferior o del extremo izquierdo

6. El jefe de operaciones de una empresa asegura que la maquinaria utilizada no está siendo aprovechada a su máxima capacidad, ya que cuando fue adquirida el proveedor aseguró que cada máquina podía producir 9 000 artículos diariamente, y una muestra aleatoria de 45 días demuestra que cada equipo produce 7 800 artículos en promedio por día con una desviación estándar de 4 300. ¿Es correcto aceptar el planteamiento del jefe de operaciones con un nivel de significación de 3 por ciento?
7. El área de mercadotecnia de una tienda de autoservicio tiene registrado que el trimestre antepasado los clientes gastaron en promedio $121 en sus compras. Se quiere saber si en el trimestre pasado dicha cifra descendió, así que toman una muestra aleatoria de 90 compras del último trimestre y encuentran que en promedio los clientes consumieron $114 por compra con una desviación estándar de $42.6. Con un nivel de significación de 6% ¿puede asegurarse que este promedio disminuyó?
8. Un laboratorio acaba de mejorar una de sus fórmulas para que el tiempo de reacción promedio en el organismo sea menor. Con una muestra aleatoria de 32 pacientes se encuentra que el promedio es de 7 días con una desviación estándar de 2.35, mientras que normalmente es de 8 días. Con un nivel de significación de 0.08, ¿es posible asegurar que el tiempo promedio de reacción disminuyó?

Prueba de hipótesis de la cola superior o del extremo derecho

9. Una empresa de seguros ha estado aplicando diferentes técnicas para incrementar sus ventas durante los últimos 6 meses. El promedio de ventas por semestre es de 54 ventas diarias; con una muestra aleatoria de 60 días de los últimos 6 meses, se obtiene que en promedio hay 60 ventas diarias con una desviación estándar de 28. Con un nivel de significación de 5%, ¿es posible asegurar que el promedio de ventas aumentó?
10. En una fábrica de motores se desea reducir el tiempo de ensamblado y se implementa un nuevo proceso para lograr ese fin. Normalmente ensamblar esos motores tomaba en promedio 12 minutos. Con el nuevo proceso, una muestra aleatoria de 42 motores refleja que el ensamblado tarda 15 minutos con una desviación estándar de 11 minutos. Con un nivel de significación de 4%, ¿podemos asegurar que el tiempo promedio de ensamblaje aumentó?
11. Un banco desea saber si las transacciones promedio que realizan sus clientes diariamente superan los 500 000 pesos, entonces se toma una muestra aleatoria de 35 días y se encuentra que el promedio del monto de las transacciones diarias es de 510 000 pesos con una desviación estándar de 29 000 pesos. ¿Es posible asegurar que este monto incrementó con un nivel de significación de 2 por ciento?

9.7 Métodos para realizar pruebas de hipótesis

Método del intervalo. Construcción de un intervalo que contenga el valor muestral observado de acuerdo con el valor de la media poblacional planteado en la hipótesis nula y el nivel de significación.

De los ejemplos 9.1 al 9.3 se ilustró la realización de una prueba de hipótesis utilizando el **método del intervalo** que consiste en construir un intervalo dentro del cual se espera que esté el valor muestral observado de acuerdo con:

a) el valor de la media poblacional planteado en la hipótesis nula y
b) el nivel de significación.

En esta sección se repasarán estos 3 ejemplos y se explicarán otros 2 métodos: el del valor crítico del estadístico de prueba y el del valor de la P (la probabilidad).

Se ilustran los otros 2 métodos con los mismos datos para enfatizar que se trata de métodos completamente equivalentes; se llega a los mismos resultados por caminos distintos.

En términos generales, el método del intervalo se desprende directamente de la metodología para estimar parámetros utilizada en el capítulo anterior; este método como los otros 2 que se estudian en esta sección son equivalentes, ya que simplemente utilizan la misma información para llegar, por diferente camino, a los mismos resultados.

Vale la pena anotar que el método del intervalo que se aplicó inicialmente es útil como introducción a la metodología de las pruebas de hipótesis por 2 razones principales:

1. Se desprende directamente de la metodología de la estimación de parámetros que se estudió en el capítulo anterior.
2. Porque permite abundar en la relación entre la probabilidad especificada mediante el nivel de significación α, el valor de z que corresponde a una α determinada y entre esa misma probabilidad expresada en términos de la variable original en la que se plantean los casos a resolver.

Consideramos que es muy importante que el estudiante comprenda cabalmente estas relaciones.

Con respecto a los 2 nuevos métodos que se presentan ahora, conviene mencionar el origen de su importancia; son relevantes porque al conocer los 3 métodos se permite una mejor comprensión de los razonamientos que subyacen a las pruebas de hipótesis, que a su vez es la técnica más importante de la inferencia estadística. En lo particular, el método del valor crítico del estadístico de prueba es de uso común porque existen muchos ensayos de hipótesis en donde se utilizan distribuciones de probabilidad distintas a la normal y su z correspondiente. Por lo revisado en el capítulo anterior se puede visualizar que cuando las circunstancias lo ameritan (cuando el tamaño de la muestra es menor de 30, cuando se desconoce la desviación estándar de la población y cuando la variable se distribuye normalmente) el estadístico de prueba apropiado es la t de Student. De la misma manera en otros capítulos se verán pruebas de hipótesis que utilizan los estadísticos de prueba correspondientes a diferentes distribuciones de probabilidad.

Por su parte el método del valor de la P, la probabilidad, es un tercer método que se desarrolla directamente con el nivel de significación, y es un método utilizado ampliamente en el medio de las publicaciones científicas para la presentación de resultados.

9.7.1 Método del intervalo

En este método se construyó un intervalo dentro del cual se encuentran todas las medias posibles de la distribución muestral, y la regla de decisión consiste en aceptar H_0 si el valor observado de la media en la muestra cae dentro de ese intervalo delimitado por la región (o regiones) de aceptación, y en rechazar la hipótesis nula si cae en la región de rechazo. En las 2 secciones siguientes se aplicarán los métodos estadístico de prueba y el del valor de la P.

9.7.2 Método del estadístico de prueba

Este método implica los mismos razonamientos que se siguieron en los ejemplos 9.1 al 9.3, salvo que ahora se contemplan desde el punto de vista del estadístico de prueba que en esos casos fue z el parámetro de la distribución normal estándar. Se resuelve ahora como ejemplo 9.4 el ejemplo 9.1.

EJEMPLO 9.4

Un fabricante de llantas para automóvil afirma que la duración promedio de determinado modelo de llanta es de 40 000 km bajo condiciones normales de manejo en un automóvil de cierto peso. Se analiza una muestra aleatoria de 100 llantas de este tipo, bajo

las condiciones especificadas, y se encuentra que la duración promedio fue de 39 000 km con una desviación estándar de 8 500 km. Pruebe la afirmación del fabricante con un nivel de significación de 0.05.

Solución: En primer lugar las hipótesis son las mismas:

$$H_0: \mu = 40\,000$$
$$H_1: \mu \neq 40\,000$$

Se encontró que el error estándar de la media es:

$$s_{\bar{X}} = \frac{s}{\sqrt{n}} = \frac{8\,500}{\sqrt{100}} = \frac{8\,500}{10} = 850$$

También se descubrió que los valores de z que definen las áreas de aceptación y de rechazo de acuerdo con el nivel de significación especificado de $\alpha = 0.05$ es:

$$P\,(-1.96 \leq z \leq 1.96) = 0.05$$

Por lo que el valor crítico del estadístico de prueba es, precisamente, 1.96 tanto positivo como negativo porque se trata de una prueba de 2 colas.

Si se utiliza z como el valor de referencia para tomar la decisión se determina su valor de la forma acostumbrada:

$$z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \frac{39\,000 - 40\,000}{\frac{8\,500}{\sqrt{100}}} = \frac{-1\,000}{850} = -1.176$$

Es importante recordar que este valor de –1.176 significa que la media observada de 39 000, está a 1.176 desviaciones estándar (errores estándar de la media) a la izquierda de la media según la hipótesis nula de 40 000. Y como este valor observado de –1.176 se encuentra entre los 2 valores de z que aíslan 5% del área de la curva normal en ambos extremos de la distribución, $z = -1.96$ y $z = 1.96$, entonces no se rechaza la hipótesis nula para concluir que la afirmación del fabricante sí tiene sustento, con una probabilidad de 0.05 de estar cometiendo el error de rechazar una hipótesis verdadera. Estos argumentos se presentan en la figura 9.9.

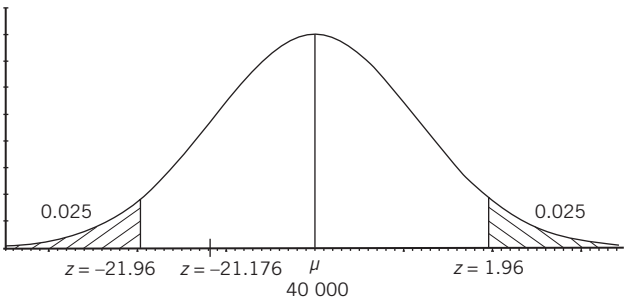


Figura 9.9 Datos del ejemplo 9.4 con el método del estadístico de prueba.

Tal como puede verse en esta figura, la región de rechazo son los 2 extremos y la región de aceptación es la parte central que contiene 95% de los casos posibles de acuerdo con H_0. Vale la pena tener en cuenta 2 detalles sobre esta división del área de la curva normal en regiones de aceptación y de rechazo:

1. Estas regiones resumen las reglas para tomar la decisión de aceptar o rechazar H_0.
2. En el caso del método del estadístico de prueba, el eje horizontal se mide en unidades de z, en tanto que en el método del intervalo, que se ilustró inicialmente, este mismo eje horizontal se mide en unidades de la variable original, km.

Se resuelve en seguida como ejemplo 9.5 el ejemplo 9.2 pero ahora con el método del estadístico de prueba.

EJEMPLO 9.5

Fumar cigarros de la marca X produce en promedio 0.6 mg de nicotina. El departamento de ingeniería del fabricante propone un filtro nuevo que supuestamente reducirá la producción de nicotina. Se toma una muestra de 50 cigarros con el nuevo filtro y se encuentra que el promedio de nicotina es de 0.55 mg, con desviación estándar de 0.56. ¿Debe aceptarse la aseveración del departamento de ingeniería con un nivel de significación de 2.5 por ciento?

Solución: Las hipótesis:

$$H_0: \mu = 0.60$$
$$H_1: \mu < 0.60$$

El valor de z que divide 2.5% del área de la curva a la izquierda es –1.96; este nivel de significación en símbolos:

$$P(z \leq -1.96) = 0.025$$

Donde el valor crítico del estadístico de prueba z es –1.96, y es el valor contra el que se va a comparar el valor de z a calcular a partir de los datos muestrales:

$$z = \frac{X - \mu}{s_{\bar{X}}} = \frac{X - \mu}{\frac{s}{\sqrt{n}}} = \frac{0.55 - 0.6}{\frac{0.56}{\sqrt{50}}} = \frac{-0.05}{0.0792} = -0.6313$$

Como el valor calculado de z cae en la región de aceptación, puesto que está a la derecha del –1.96, no se rechaza la hipótesis nula y se concluye que los nuevos filtros no reducen la nicotina que producen esos cigarros; la misma conclusión a la que se llegó con el método del intervalo. En la figura 9.10 se ilustra esta información.

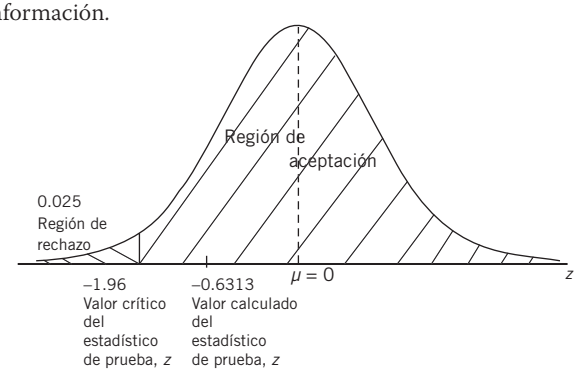


Figura 9.10 Datos del ejemplo 9.5.

Se resuelve en seguida el ejemplo 9.6 con el método del valor crítico del estadístico de prueba.

EJEMPLO 9.6

Una institución de enseñanza de idiomas registra en sus cursos a 4 personas en promedio por semana, por lo que hace una campaña publicitaria de gran magnitud. Después de un año se toma una muestra aleatoria de 32 semanas y se encuentra que en promedio se registran semanalmente 5 alumnos nuevos con una desviación estándar de 0.9. Tomando en cuenta un nivel de significación de 0.05, determine si la campaña publicitaria cumplió su objetivo.

Solución: Las hipótesis:

$$H_0: \mu = 4$$
$$H_1: \mu > 4$$

La z crítica es 1.645 ya que $P(z \geq 1.645) = 0.05$, que es el nivel de significación.

La z calculada con los datos muestrales:

$$z = \frac{X - \mu}{s_{\bar{X}}} = \frac{X - \mu}{\frac{s}{\sqrt{n}}} = \frac{5 - 4}{\frac{0.9}{\sqrt{32}}} = \frac{1}{0.159} = 6.29$$

Como el valor calculado de la z cae en la región de rechazo (es mucho mayor que el valor crítico), se refuta H_0 y se concluye, igual que antes, que la campaña publicitaria cumplió con su objetivo. En la figura 9.11 se ilustra esta información.

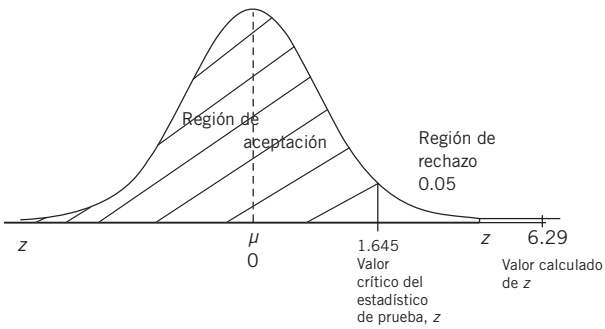


Figura 9.11 Los datos del ejemplo 9.6.

9.7.3 Método del valor de la *P*

Método del valor de la *P*. Ayuda a comprender los procedimientos de prueba de hipótesis; reporta conclusiones sobre estudios estadísticos en publicaciones científicas.

Una tercera manera de interpretar los resultados obtenidos de una muestra para decidir si se acepta o no tal o cual hipótesis es lo que se conoce como el **método del valor de la *P***. Este método, además de ayudar a complementar la comprensión de los procedimientos de prueba de hipótesis, es una forma que se acostumbra para reportar conclusiones sobre diversos estudios estadísticos en publicaciones científicas.

El método consiste, básicamente, en determinar la probabilidad P de haber obtenido el valor observado, la media de la muestra en el ejemplo, asumiendo que la hipótesis nula es cierta. Para tomar la decisión de aceptar o rechazar esta H_0 se compara este valor de P con α, el valor especificado del nivel de significación. Si P es menor que α se rechaza H_0; si es mayor se acepta.

EJEMPLO 9.7

De nueva cuenta el ejemplo de las llantas, en el que se vio que:

$$z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \frac{39\,000 - 40\,000}{\frac{8\,500}{\sqrt{100}}} = \frac{-1\,000}{850} = -1.176$$

También se reconoció que este valor de –1.176 significa que la media observada, 39 000, está a 1.176 desviaciones estándar (errores estándar de la media) a la izquierda de la media según la hipótesis nula, 40 000.

De la tabla de áreas bajo la curva normal, $P(z \leq -1.176) = 0.1198$. Y, como esta probabilidad es mayor que el nivel de significación dividido entre 2 puesto que se trata de una prueba de dos colas, $0.05/2 = 0.025$, no se rechaza H_0; se concluye, igual que antes, que el fabricante tiene razón cuando afirma que el promedio de duración de sus llantas es de 40 000 km.

En otras palabras, esta P dice que si la hipótesis nula es cierta existe una probabilidad de 0.119 u 11.9% de haber obtenido una muestra aleatoria con una media de 40 000 km, y esta probabilidad es lo suficientemente grande en comparación con el nivel de significación de 5% como para rechazar H_0, por lo que se le acepta. En la figura 9.12 se ilustra esta información.

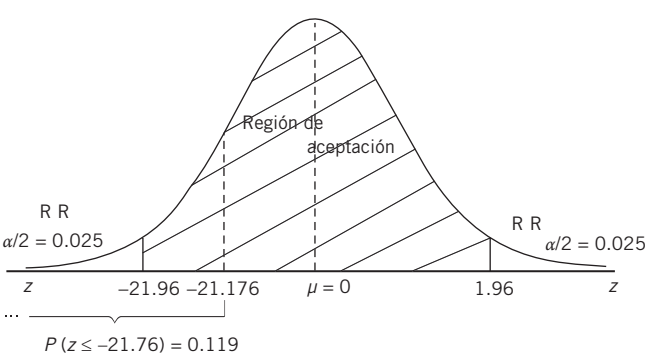


Figura 9.12 Datos del ejemplo 9.7.

En los 2 ejemplos siguientes se resuelven, con el método del valor de la P los ejemplos 9.2 y 9.3, respectivamente, que fueron resueltos con los otros 2 métodos.

Ejemplo 9.8

En el caso expuesto sobre la nicotina que producen ciertos cigarros se tenían las siguientes hipótesis:

$$H_0: \mu = 0.60$$
$$H_1: \mu < 0.60$$

Con un nivel de significación, α, de 0.025 o 2.5% el valor calculado de la z fue:

$$z = \frac{X - \mu}{s_{\bar{X}}} = \frac{X - \mu}{\frac{s}{\sqrt{n}}} = \frac{0.55 - 0.6}{\frac{0.56}{\sqrt{50}}} = \frac{-0.05}{0.0792} = -0.6313$$

Y de la tabla de áreas bajo la curva normal tenemos que:

$$P(z < -0.6313) = 0.2639$$

En la figura 9.13 se ilustra esta información.

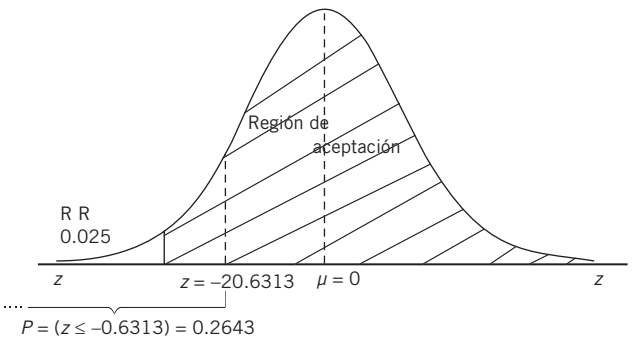


Figura 9.13 Datos del ejemplo 9.8.

Debido a que esta probabilidad es mayor que el nivel de significación de 0.025 no se rechaza la hipótesis nula, al igual que antes.

Ejemplo 9.9

En el caso de los alumnos que se registran en la escuela de idiomas se tenían las siguientes hipótesis:

$$H_0: \mu = 4$$
$$H_0: \mu > 4$$

Además, el valor calculado de la z con los valores muestrales fue de:

$$z = \frac{X - \mu}{s_{\bar{X}}} = \frac{X - \mu}{\frac{s}{\sqrt{n}}} = \frac{5 - 4}{\frac{0.9}{\sqrt{32}}} = \frac{1}{0.159} = 6.29$$

Aun sin consultar la tabla de áreas bajo la curva normal se sabe que la probabilidad de haber obtenido este valor siendo H_0 cierta es prácticamente de cero. Por lo que, siendo el nivel de significación de 0.05, se rechaza la hipótesis nula para concluir que la campaña publicitaria sí aumentó la inscripción de alumnos.

En la figura 9.14 se ilustra esta información.

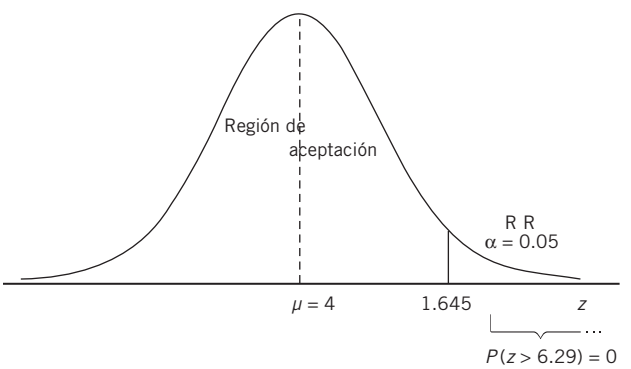


Figura 9.14 Datos del ejemplo 9.9.

En lo sucesivo se utilizará principalmente el método del estadístico de prueba porque es el más adecuado para este texto, ya que es especialmente útil para ilustrar el manejo de diferentes estadísticos de prueba. Sin embargo no se dejará de lado a los otros 2 métodos porque resultan eficaces y convenientes.

9.7.4 Resumen de los procedimientos para realizar pruebas de hipótesis con los 3 métodos

Destacan las actividades comunes a los 3 métodos:

1. Formular las hipótesis nula y alternativa, según el planteamiento del caso.
2. Determinar α, el nivel de significación.

3. Elaborar una gráfica que facilite la visualización de las condiciones de la prueba.
4. Identificar las regiones de aceptación y de rechazo, según lo indiquen las hipótesis.
5. Calcular el error estándar del estadístico que, como se verá más adelante, incluye determinar inicialmente cuál será el estadístico de prueba. Como hasta aquí sólo se ha utilizado z, el cálculo es, en su forma reducida (sin el factor de corrección por población finita):

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}}$$

Cuando se conoce σ la desviación estándar de la población, o

$$s_{\bar{X}} = \frac{s}{\sqrt{n}}$$

Si se desconoce σ, y se utiliza la desviación estándar de la muestra.

6. Con los 3 métodos

6.1. Para el método del intervalo:

a) Construir éste como:

$\mu \pm z\sigma_{\bar{X}}$ para una prueba de dos colas
$\mu + z\sigma_{\bar{X}}$ para una prueba de la cola superior, y
$\mu - z\sigma_{\bar{X}}$ para una prueba de la cola inferior.

b) Determinar si el valor observado de la media en la muestra se encuentra dentro de la zona de rechazo o en la de aceptación, decidir si se rechaza o no la hipótesis nula y, en consecuencia, si se acepta o no la hipótesis alternativa.

c) Interpretar la decisión en términos del problema planteado.

6.2. Para el método del valor crítico del estadístico de prueba:

a) Determinar el valor crítico del estadístico de prueba con relación al nivel de significación establecido y de acuerdo con su condición de prueba de 1 o de 2 extremos o colas. En los casos que se revisaron, este valor crítico se determina a través de la tabla de áreas bajo la curva normal.

b) Calcular el valor muestral del estadístico de prueba, de la siguiente manera:

$$Z = \frac{X - \mu}{\sigma_{\bar{X}}} = \frac{X - \mu}{\frac{\sigma}{\sqrt{N}}}$$

Si se calcula el error estándar de la media a partir de la desviación estándar de la población, y

$$Z = \frac{X - \mu}{s_{\bar{X}}} = \frac{X - \mu}{\frac{s}{\sqrt{n}}}$$

En caso de que se calcule el error estándar de la media a partir de datos muestrales:

c) Determinar si el valor de z observado en la muestra se encuentra en la zona de rechazo o en la de aceptación (establecida con respecto a la z crítica). Decidir si se rechaza o no la hipótesis nula y, en consecuencia, si se acepta o no la hipótesis alternativa.

d) Interpretar la decisión en términos del problema planteado.

6.3. Para el método del valor de la P:

a) Determinar la probabilidad de que la media observada en la muestra pudiera haberse obtenido en las condiciones planteadas, calculando el valor de z con los datos muestrales, tal como se especifica en el inciso *b*) anterior.

Una vez obtenido el valor de z con los datos muestrales se utiliza la tabla de áreas bajo la curva normal para determinar la probabilidad de haberlo obtenido, que es precisamente la P, y se compara esa probabilidad con α, el nivel de significación. Como se mencionó, si P es menor que α se rechaza H_0; si es mayor, no.

EJERCICIOS 9.7 Métodos para realizar pruebas de hipótesis

Resuelva los ejercicios 1 a 11 de la sección 9.6 que fueron resueltos por los métodos del intervalo pero ahora utilice los métodos del estadístico de prueba y el de la P, la probabilidad.

9.8 Prueba de hipótesis sobre una proporción poblacional

En lo revisado hasta aquí, se han resuelto casos de pruebas de hipótesis sobre medias (promedios) poblacionales. En esta sección se ilustran los procedimientos estadísticos aplicables para probar hipótesis sobre proporciones poblacionales y se utilizan los 3 métodos.

El proceso que debe llevarse a cabo para realizar pruebas de hipótesis con proporciones es prácticamente igual al que se siguió para promedios. En el caso de proporciones deben utilizarse los valores correspondientes a esta medida. En seguida se ilustra con algunos ejemplos.

EJEMPLO 9.10

En China, un fabricante de juguetes afirma que sólo 10% o menos del total de osos de peluche parlantes que produce están defectuosos. Se sometieron a prueba en forma aleatoria a 400 de estos juguetes y se encontró que 50 estaban defectuosos. Compruebe la afirmación del fabricante con un nivel de significación de 5 por ciento.

Solución:

1. En primer lugar, las hipótesis:

$$H_0: \pi \leq 0.10$$
$$H_1: \pi > 0.10$$

2. El nivel de significación:

$$\alpha = 0.05$$

3 y 4. Debido a que las hipótesis implican desigualdad se trata de una prueba de un extremo como se ilustra en la figura 9.5, donde se señalan las regiones de aceptación y de rechazo, de acuerdo con el nivel de significación α.

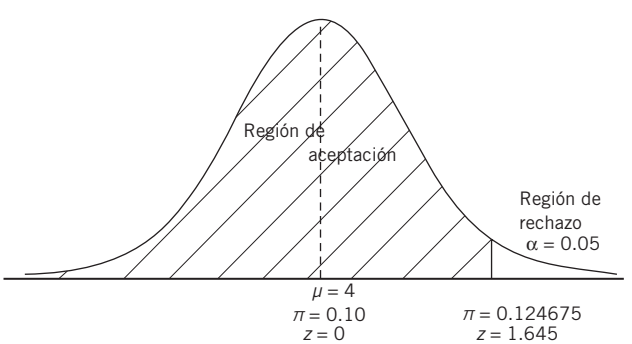


Figura 9.15 Distribución normal con los datos del ejemplo 9.10.

5. Tal como se vio en el capítulo anterior, el error estándar de la proporción es:

$$\sigma_p = \sqrt{\frac{\pi Q}{n}} = \sqrt{\frac{0.1(0.9)}{400}} = \sqrt{0.000225} = 0.015$$

6. Con los 3 métodos.

6.1. Con el método del intervalo se obtiene que éste es:

$$\pi + 1.645\ (0.015) = 0.10 + 0.024675 = 0.124675$$

De tal suerte que el intervalo es:

$$-\infty \leq p \leq 0.124675$$

Este valor del límite superior del intervalo ya se había anotado en la gráfica anterior.

Ahora, en razón de que la proporción observada en la muestra es $p = \frac{50}{400} = 0.125$ y sale del intervalo establecido se rechaza H_0; se concluye que la proporción de osos parlantes defectuosos fabricados por ese empresario chino es mayor de 10 por ciento.

6.2. Con el método del valor crítico del estadístico de prueba: el valor crítico de z que divide la región de aceptación de la de rechazo es $z \geq 1.645$, ya que $P(z \geq 1.645) = 0.05$. Ahora, el valor calculado de z

$$z = \frac{p - \pi}{\sigma_p} = \frac{p - \pi}{\sqrt{\frac{\pi Q}{n}}} = \frac{0.125 - 0.10}{\sqrt{\frac{0.1(0.9)}{400}}} = \frac{0.025}{0.015} = 1.67$$

Como 1.67 es mayor que el crítico 1.645, se rechaza H_0, de la misma manera que anteriormente se hizo. Se concluye que la proporción de osos parlantes defectuosos no es de 10%, sino mayor.

6.3. Con el método de la P:

La probabilidad de obtener un valor de z de 1.67 o mayor es de 0.0475; esta probabilidad es menor que el nivel de significación $\alpha = 0.05$, por lo que se concluye que puede rechazarse la hipótesis nula y concluir que la proporción de osos parlantes defectuosos no es de 10% sino mayor.

Ejemplo 9.11

Con base en estudios anteriores, se sabe que la proporción laboralmente activa de los estudiantes de una universidad es de 30%. Se desea probar si esta información sigue siendo válida y se toma una muestra aleatoria de 200 estudiantes donde se descubre que 70 de ellos trabajan. ¿Puede afirmarse que la proporción de estudiantes trabajadores sigue siendo de 30%? Probar la hipótesis con un nivel de significación de 1 por ciento.

Solución:

1. Las hipótesis:

$$H_0: \pi = 0.30$$
$$H_1: \pi \neq 0.30$$

2, 3 y 4. Dado el planteamiento que conduce a estas hipótesis se reconoce que es una prueba de 2 extremos y la probabilidad α de 0.01 se divide entre los 2 extremos de la curva normal; se tiene que los valores de z que dividen las regiones de aceptación y de rechazo son: $-2.575 \leq z \leq 2.575$. En la figura 9.16 se ilustra esta información.

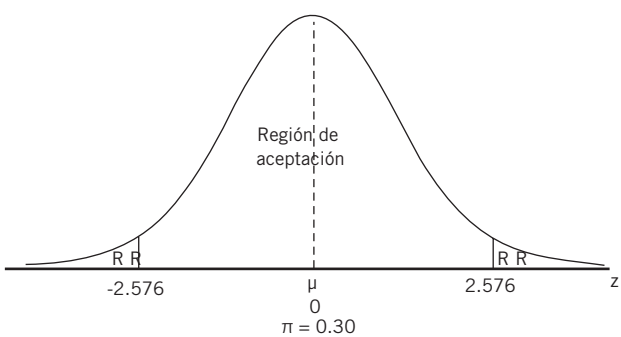


Figura 9.16 Datos ilustrados del ejemplo 9.11.

5. El error estándar de la proporción:

$$\sigma_p = \sqrt{\frac{\pi Q}{n}} = \sqrt{\frac{0.3(0.7)}{200}} = \sqrt{0.00105} = 0.0324$$

6. Con los 3 métodos:

6.1. Con el método del intervalo y estos datos, el intervalo es:

$$\pi \pm 2.575\ (0.0324) = 0.30 \pm 0.0834$$
$$= 0.2166 \text{ a } 0.3834$$

De tal suerte que el intervalo es: $0.2166 \leq p \leq 0.3834$.

Ahora, como la proporción observada en la muestra es: $p = \frac{70}{200} = 0.35$ y cae dentro del intervalo establecido no se rechaza H_0; se concluye que la proporción laboralmente activa de estudiantes de esta universidad sigue siendo de 30 por ciento.

6.2. Con estos datos y el método del valor crítico del estadístico de prueba.

La proporción de la muestra es $p = \frac{70}{200} = 0.35$, por lo que:

$$z = \frac{p - \pi}{\sigma_p} = \frac{p - \pi}{\sqrt{\frac{\pi Q}{n}}} = \frac{0.35 - 0.30}{0.0324} = \frac{0.05}{0.015} = 1.54$$

Ya que este valor del estadístico de prueba, calculado con los datos muestrales, es mayor que $z = -2.575$ y menor que $z = 2.575$, no es posible rechazar la hipótesis nula y se concluye que la proporción de estudiantes de esa universidad que trabajan sigue siendo 30 por ciento.

6.3. Con esos mismos datos y el método de la P.

Después de que se ha calculado el valor de z con los datos muestrales se determina que la probabilidad de que z sea mayor que 1.54 es de $P(z > 1.54) = 0.0618$, o sea, 6.18%. Como esta probabilidad es mayor que el nivel de significación de 0.005 o 0.5% del extremo derecho, se llega a la misma conclusión: no es posible rechazar la hipótesis nula y se averigua que la proporción de estudiantes de esa universidad que trabajan sigue siendo de 30 por ciento.

Ejemplo 9.12

El coordinador de la bolsa de trabajo de una universidad pública afirma que cuando menos 30% de los alumnos que terminan sus estudios obtiene empleo antes de 3 meses. Para probar esta afirmación se toma una muestra de 50 estudiantes de dicha institución y se encuentra que sólo 10 se emplearon durante los 3 meses después de haber finalizado sus estudios. ¿Puede rechazarse la afirmación del coordinador con un nivel de significación de 1 por ciento?

Solución: Al tratarse de una muestra grande puede utilizarse z como estadístico de prueba.

1. Las hipótesis:

$$H_0: \pi = 0.30$$
$$H_1: \pi \neq 0.30$$

2, 3 y 4. El planteamiento que conduce a estas hipótesis refiere que se trata de una prueba de 2 extremos; como $\alpha = 0.01$,

se divide esta probabilidad entre los 2 extremos de la curva normal y se tiene que los valores de z que dividen las regiones de aceptación y de rechazo son: $-2.575 \leq z \leq 2.575$, ya que $P(-2.575 \leq z \leq 2.575) = 0.99$. En la figura 9.17 se ilustra esta información.

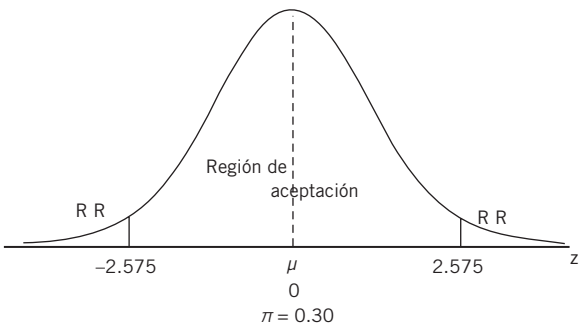


Figura 9.17 Regiones de aceptación y de rechazo para el ejemplo 9.12.

5. El error estándar de la proporción:

$$\sigma_p = \sqrt{\frac{\pi Q}{n}} = \sqrt{\frac{0.3(0.7)}{50}} = \sqrt{0.0042} = 0.065$$

La proporción de la muestra es $p = 10/50 = 0.20$, por lo que:

$$z = \frac{p - \pi}{\sigma_p} = \frac{p - \pi}{\sqrt{\frac{\pi Q}{n}}} = \frac{0.20 - 0.30}{0.065} = \frac{0.10}{0.065} = 1.54$$

Ya que este valor del estadístico de prueba calculado con los datos muestrales es mayor que $z = -2.576$ y menor que $z = 2.576$, no es posible rechazar la hipótesis nula y se concluye que la proporción de estudiantes que terminan sus estudios y que consiguen empleo antes de 3 meses sigue siendo 30 por ciento.

ejercicios 9.8 Prueba de hipótesis sobre una proporción poblacional

Pruebas de hipótesis para proporción

De 2 colas

1. Una mueblería ha tenido muchos problemas para vender determinado producto en los últimos 3 meses, por esta razón ofrece promociones para que sus clientes lo compren. La venta de este producto sólo representa 15% de sus ventas totales, el gerente del lugar supone que en 2 meses este porcentaje cambió, por lo que emplea una muestra aleatoria de las ventas de 45 días y encuentra que de cada 100 compras 22 son de este producto. Determine si la suposición es cierta o falsa y en caso de ser verdadera, aclarar si el cambio fue bueno o malo, tomando en cuenta un nivel de significación de 0.04.
2. Para tener mayor control en las carreteras, el gobierno decidió realizar retenes en las salidas de la ciudad para observar qué porcentaje de conductores contaban con la documentación necesaria para transitar. En los primeros 6 meses se registró que de cada 25 conductores 17 cumplieron con todos los requisitos. Después de este tiempo de prueba, las autoridades suponen que debe haber un cambio significativo en el porcentaje, por lo que toman una muestra aleatoria de 120 conductores y encuentran que 94 de ellos presentaron todos sus documentos. Determine si hubo un cambio en el porcentaje con un nivel de significación de 2 por ciento.

Cola inferior, lado izquierdo

3. Una fábrica de refacciones automotrices produce diariamente 1 230 piezas, de las cuales 65 resultan defectuosas. Con el paso del tiempo, el supervisor de la producción supuso que el defecto de estos productos podía evitarse si se realizaba un insignificante procedimiento; así que desde hace 2 meses los trabajadores de la planta lo llevan a cabo. Con una muestra aleatoria de 37 productos, se obtiene que el porcentaje de productos defectuosos es 15%. Determine con un nivel de significación de 3% si el porcentaje realmente disminuyó.
4. El área de recursos humanos de una empresa tiene registrado que el año pasado 55% de los trabajadores deseaban cambiarse del área laboral a la que pertenecían. Este año se hizo un gran número de cambios de personal entre las áreas, por lo que se espera que aquel porcentaje haya disminuido. Se toma una muestra aleatoria de 90 empleados y se encuentra que 40 continúan en la posición de querer cambiar de área. Determine si realmente menos trabajadores quieren cambiarse de área a comparación del año pasado, con un nivel de significación de 0.05.

Cola superior, lado derecho

5. Un banco encuentra que sólo 25% de sus clientes tienen cuentas con saldos superiores a 30 000 pesos, entonces disminuye las tasas de interés y ofrece mayores beneficios para incrementar la cantidad de clientes con esta característica. Una vez pasados seis meses se toma una muestra aleatoria de 250 clientes y se obtiene que 83 de ellos tienen cuentas con un saldo que rebasa los 30 000 pesos. Si se tiene un nivel de significación de 1%, determine si realmente este porcentaje se incrementó.
6. El director de una universidad encuentra que únicamente 30% de los alumnos toman cursos extracurriculares y piensa que se debe a los horarios, por lo que realiza un estudio sobre ellos y hace algunos cambios. Con una muestra aleatoria de 400 alumnos encuentra que 32% de ellos toman estos cursos, y desea saber con un nivel de significación de 0.05, si realmente el porcentaje se incrementó.

9.9 Resumen

En este capítulo se introdujeron los conceptos básicos sobre uno de los 2 principales métodos de la inferencia estadística: las pruebas de hipótesis.

Una hipótesis estadística es una suposición sobre el valor de una medida de población, como una media o una proporción. Asimismo se explicó que el paso inicial de las pruebas de hipótesis consiste en plantear la *hipótesis nula*, que es la suposición a evaluar o probar, junto con una hipótesis alternativa, la que resulta ser aceptable como cierta en caso de que la nula no lo sea.

El mecanismo muestral para decidir si una hipótesis nula es cierta o no consiste en analizar si los datos obtenidos de la muestra son consistentes con ella o no. Es claro que si estos datos muestrales son consistentes con la hipótesis nula se concluye que no hay elementos para rechazarla, por lo tanto es cierta. Ahora bien, si los datos muestrales no son consistentes con la hipótesis nula, como no la apoyan, hay elementos para concluir que no es cierta y se rechaza. Entonces, la hipótesis alternativa es la cierta.

Se explicaron los 2 tipos de errores que pueden cometerse al realizar una prueba de hipótesis. Por un lado está el rechazo de una hipótesis cierta (error tipo I), por el otro aceptar una falsa (error tipo II). Del mismo modo, se estableció que se denomina como *nivel de significación* a la máxima probabilidad que se acepta de cometer un error tipo I en una prueba de hipótesis y se le identifica como α. Por su parte, a la probabilidad de cometer un error de tipo II se le identifica como β.

Se sugirió que es una buena idea elaborar una gráfica de las condiciones de una prueba de hipótesis para visualizar mejor las circunstancias y los planteamientos. Se propone el siguiente procedimiento para realizar estas pruebas:

1. Plantear la hipótesis nula y la alternativa.
2. Establecer el nivel de significación al que se desea realizar la prueba.
3. Calcular el error estándar del estadístico (en el ejemplo, el error estándar de la media).
4. Con base en ese nivel de significación y en el error estándar del estadístico, establecer la regla de decisión, es decir, los valores para los que se acepta la hipótesis nula y aquellos para los cuales se rechaza. Como se verá más adelante, estos valores que permiten aceptar o rechazar la hipótesis nula (y en contraparte aceptar o rechazar la alternativa), dependen de cuál de los 3 métodos de prueba se aplica: *a*) intervalo, *b*) estadístico de prueba o *c*) *p*, la probabilidad.
5. Tomar la decisión con base en todos los elementos anteriores.
6. Finalmente, de suma importancia, interpretar los resultados en términos del planteamiento original.

Se explicaron los 3 tipos de casos que se presentan en las pruebas de hipótesis: pruebas de hipótesis de 2 extremos o colas, que implican hipótesis de igual o no igual (= y ≠); pruebas del extremo izquierdo y del extremo derecho que implican hipótesis de igual y de menor o mayor que (=, <, y >).

Finalmente, se ilustraron los procedimientos para probar hipótesis sobre una media y sobre una proporción poblacionales; también se ejemplificaron los 3 métodos mediante los cuales se pueden llevar a cabo las pruebas:

1. El método del intervalo.
2. El método del estadístico de prueba.
3. El método de la *P*, la probabilidad.

9.10 Excel Uso de Excel

La única función que ofrece Excel para realizar pruebas de hipótesis es la que se llama *prueba.z*, ésta permite calcular el valor de probabilidad de una cola de una prueba con la distribución normal para una media aritmética. En otras palabras, esta función utiliza el método de la *P* para realizar la prueba y su sintaxis es:

PRUEBA.Z(matriz,μ_0,sigma)

En donde:

Matriz es el conjunto de datos, es decir, el conjunto de observaciones de la muestra con el que debe realizarse la prueba. Se especifica mediante el conjunto de celdas de la hoja de Excel que los contiene. μ_0 es el valor que se supone en la hipótesis nula para la media poblacional.

Sigma es la desviación estándar de la población, que se asume conocida. Si se omite se utiliza la desviación estándar de la muestra.

Con esta información puede comprenderse que la prueba se realiza sobre el conjunto de datos muestrales, en este caso se ahorraría el cálculo de la media y la desviación estándar muestrales. Sin embargo, en este capítulo se ha orientado el análisis con mayor tendencia hacia la metodología de las pruebas de hipótesis porque se asume que ya se conocen los procedimientos para calcular medias y desviaciones estándar, de tal manera que lo importante aquí son los procedimientos para realizar las pruebas.

Por lo anterior, no se abunda más sobre esta prueba de Excel dado lo reducido de su aplicabilidad. No sólo se reduce a medias aritméticas, sino que impone conocer todos los datos muestrales y además es únicamente para pruebas de un solo extremo.

9.11 Ejercicios adicionales

9.6.1 Pruebas de 2 extremos o colas

1. Una fábrica de refacciones invirtió en una máquina que produciría 3 mil piezas en una hora, según lo acordado con el vendedor. El jefe de producción toma una muestra aleatoria de 48 horas de trabajo de la máquina y encuentra que en promedio ha producido 2 970 piezas con una desviación estándar de 22. Compruebe la afirmación del vendedor de la máquina con un nivel de significación de 0.05.
2. La duración de los focos de 100 W que comercializa una empresa es de 750 horas según el fabricante. Para comprobarlo, el gerente de ventas toma una muestra aleatoria de 150 focos y encuentra que en promedio duran 730 horas con una desviación estándar de 110 horas. ¿Puede asegurarse que la afirmación del fabricante es cierta, tomando en cuenta un nivel de significación del 0.02?
3. En una avenida se señala que los autos deben transitar a una velocidad máxima promedio de 60 km/h. Una muestra aleatoria de 81 autos refleja que la velocidad máxima promedio es de 64.3 km/h con una desviación estándar de 7 km/h. Con un nivel de significación de 0.01, determine si puede considerarse que se respeta el señalamiento.
4. El jefe de producción de una empresa sabe que el tiempo promedio que toma una operación es de 5 min por persona; para una evaluación de desempeño es tomada una muestra aleatoria de 300 operaciones y se encuentra que el tiempo promedio es de 4.95 min con una desviación estándar de 0.1 min. Determine si se está cumpliendo con el tiempo establecido para este proceso con un nivel de significación de 0.04.
5. Una empresa de alimentos asegura en su empaque que sus productos contienen 250 g. En una inspección se toma una muestra aleatoria de 40 productos y se encuentra que en promedio contienen 249.99 g con una desviación estándar de 0.001 g. Tomando en cuenta un nivel de significación de 0.05, compruebe la afirmación del contenido del empaque.

9.6.2 Prueba de hipótesis de la cola inferior o del extremo izquierdo

6. De acuerdo con los últimos resultados obtenidos, un restaurante tiene como dato que sus clientes consumen en promedio $185 por persona, hace 2 meses cambiaron al chef y la administración asegura que los comensales están consumiendo menos. Se toma una muestra aleatoria de 50 clientes y se encuentra que el consumo promedio por cliente es de $140 con una desviación estándar de $25. Determine con un nivel de significación de 0.01 si la administración está en lo correcto.
7. Uno de los gastos principales en una maquila es la reparación constante del equipo, ya que en promedio 12 máquinas tienen problemas diariamente, por lo que se da una capacitación al personal con el fin de disminuir ese número. A los 3 meses del curso, se toma una muestra aleatoria de 36 días y se descubre que en promedio 7 máquinas deben mandarse a reparación con una desviación estándar de 1.5. Con un nivel de significación de 0.03, determine si la capacitación cumplió su objetivo.
8. Un laboratorio desarrolló un tratamiento para disminuir peso y lo aplicó a un grupo de personas cuyo peso promedio era de 85 kg; luego de 3 semanas del inicio del tratamiento, se observa en una muestra aleatoria de 65 personas que el peso promedio es de 83.5 kg con una desviación estándar de 0.80 kg. Con un nivel de significación de 0.05, determine si la pérdida de peso es significativa.
9. El manual de operaciones de una empresa tiene como regla que el número máximo de productos defectuosos promedio diario es de 30 piezas. El personal ha sido sustituido y contemplan la posibilidad de que este promedio sea menor. Una muestra aleatoria de 35 días arroja 29.5 productos defectuosos con una desviación estándar de 2 productos. ¿Puede considerarse que disminuyó el promedio con un nivel de significación de 0.06?
10. En una delegación se implementó una campaña para disminuir el desperdicio diario de agua por persona, se sabe que en promedio se consumen 350 L. Después de realizar la campaña se tomó una muestra aleatoria de 1 500 personas y se encuentra que su consumo promedio diario es de 345 L con una desviación estándar de 13 L. Determine con un nivel de significación de 0.01 si el desperdicio de agua disminuyó.

9.6.3 Prueba de hipótesis de la cola superior o del extremo derecho

11. La directora de una escuela primaria, de acuerdo con una encuesta, sabe que en promedio sus alumnos consumen 3 veces a la semana una porción adecuada de verduras. Ofrece, a los padres de familia, una conferencia informativa sobre la importancia de la alimentación en los niños, a fin de incrementar el consumo de verduras. Después de un mes de la conferencia, se toma una muestra aleatoria de 170 niños y se obtiene que en promedio los niños consumen verduras 4 veces a la semana con una desviación estándar de 0.35. Determine si la conferencia causó el efecto deseado de acuerdo con un nivel de significación de 0.07.
12. Como medida de seguridad un banco está promoviendo el uso de tarjeta para realizar compras en cierta cadena de tiendas, en las que se tiene registro diario de 230 compras efectuadas con tarjeta. Tras la promoción se toma una muestra aleatoria de 64 días y se encuentra que el promedio de compras con tarjeta es de 245 con una desviación estándar de 28. Determine con un nivel de significación de 0.05 si hubo un aumento significativo.
13. Una empresa tiene planes de expansión y exige al personal dominar un alto nivel de inglés para seguir siendo parte de la compañía. Un examen de nivelación revela que los empleados tienen en promedio una calificación de 510 (escala de 300 a 700). A las 3 semanas se aplica nuevamente el mismo examen a la espera de obtener un mejor resultado; en una muestra aleatoria de 50 empleados se encuentra que la

calificación promedio es de 515 con una desviación estándar de 30. ¿Hubo realmente un incremento en la calificación de la evaluación de los empleados teniendo en cuenta un nivel de significación de 0.02?

14. Una distribuidora registra que en promedio hace 115 entregas diarias; un mes atrás cambió el equipo de transporte y tomó una prueba aleatoria de 36 días para comparar si el número de entregas diarias era superado con estas nuevas condiciones. Se encontró que en promedio se hacen 163 entregas con una desviación estándar de 4.7. Determine con un nivel de significación de 0.01 si el incremento es significativo.

9.7 Métodos para realizar pruebas de hipótesis:
9.7.1 Método del intervalo

15. Un profesor hace un estudio a sus alumnos y descubre que, en promedio, leen 3 libros al año (0.25 al mes), por lo que intenta recomendar buenos títulos que contribuyan al desarrollo de su curso. Pasados 6 meses, el maestro toma una muestra aleatoria de 120 estudiantes y encuentra que en promedio sus alumnos leen 0.30 libros al mes con una desviación estándar de 0.005. Determine si el incremento es significativo con un nivel de significación de 0.01.

16. Un empresario desea comprar una franquicia de helados. En el contrato se especifica que el costo promedio por producto es de $11.50, sin embargo una muestra aleatoria de 40 productos demuestra que el costo promedio es de 11.55 con una desviación estándar de 0.003. Tomando en cuenta un nivel de significación de 0.1, demuestre si los datos del contrato son verídicos.

17. Como incentivo una empresa ofrece a sus vendedores un bono si se incrementan las ventas. La condición es que el aumento sea grupal y no personal. En promedio todos los vendedores venden $230 000 a la semana. Después de un año de evaluación se toma una muestra aleatoria de 31 semanas y se observa que el promedio de ventas semanales es de $230 500 con una desviación estándar de $2 300. Tomando en cuenta un nivel de significación de 0.05, ¿es significativo el incremento de ventas?

18. Todos los días una revista recibe vía telefónica en promedio 35 sugerencias o comentarios del público, por lo que el área de mercadotecnia decide abrir para ello un espacio en la página oficial de internet. Se toma una muestra aleatoria de 60 días y se encuentra que en promedio reciben 180 participaciones del público, a través de este medio, con una desviación estándar de 17. ¿Hubo un cambio en el promedio de sugerencias o comentarios que recibe la revista diariamente? Considere un nivel de significación de 0.04.

19. Una pequeña empresa de producción de eventos realiza en promedio 3.3 proyectos diarios, 2 de los empleados más representativos renunciaron y el director cree que el número de eventos ha disminuido. Una muestra aleatoria de 50 días demuestra que en promedio se están ofreciendo 2.7 eventos diarios con una desviación estándar de 0.8. ¿Tiene razón el director? Considere un nivel de significación de 0.06.

9.7.2 Método del estadístico de prueba

20. Una embotelladora de refrescos recibe diariamente 470 envases de *pet* para reciclar; debido a que así disminuye costos y contribuye a incrementar su responsabilidad social, la empresa impulsó una campaña hace algunos meses con la intención de aumentar el retorno de envases a la empresa. Una muestra aleatoria de 100 días demuestra que regresan en promedio 485 botellas para ser recicladas con una desviación estándar de 38. Determine, con un nivel de significación de 0.01, si la campaña cumplió su objetivo.

21. El gobierno de un municipio registra mensualmente 133 denuncias por robo, razón por la cual se crea una campaña para disminuir este promedio. Pasados 3 años se toma una muestra aleatoria de 31 meses y se encuentra que en promedio se registran 129 denuncias por este delito, con una desviación estándar de 12.8. Tomando en cuenta un nivel de significación de 0.05, ¿hubo un decremento en las denuncias por robo en este municipio?

22. El área de recursos humanos de una empresa multinacional tiene como política que sus trabajadores deben recibir 20 h mensuales de capacitación constante. Una muestra aleatoria de 65 trabajadores demuestra que en promedio reciben 18 h mensuales de capacitación con una desviación estándar de 7 h. Determine, con un nivel de significación de 0.02, si en la empresa se está cumpliendo con la política establecida.

23. Una agencia que renta autos a turistas sabe que en promedio alquila 8 autos a la semana; abrieron una sucursal en otra ciudad y en una muestra aleatoria de 30 semanas se observa que en promedio rentan 11 autos a la semana con una desviación estándar de 5 autos. Teniendo en cuenta un nivel de significación de 0.07, determine si el promedio es igual en las 2 sucursales de la empresa.

24. Una diseñadora de modas sabe que en promedio los gastos de producción por pieza son de $320, sin embargo hace un tiempo el precio de las telas ha incrementado de manera significativa. Una muestra aleatoria de 32 piezas demuestra que en promedio se gastan $350 por pieza con una desviación estándar de $35. ¿Incrementaron significativamente los gastos de producción por pieza de acuerdo con un nivel de significación de 0.01?

9.7.3 Método del valor de la *P*

25. Como requisito de admisión, una universidad hace una prueba de aptitudes en la que la calificación promedio de los aspirantes es de 140 puntos. El nuevo coordinador académico quiere saber si este dato ha cambiado, por lo que toma una muestra aleatoria de 81 aspirantes y encuentra que en promedio sacan 135 puntos en la evaluación con una desviación estándar de 27 puntos. Determine, con un nivel de significación de 0.03, si hay un cambio en el promedio.

26. Una productora de materiales de construcción fabrica una baldosa que en promedio soporta 138 kg/m^2. Cambiaron de proveedor de materiales porque quieren hacer su producto

más resistente. Una prueba aleatoria de 600 baldosas de la nueva producción tiene un promedio de resistencia de 140 kg/m^2 con una desviación estándar de 20 kg/m^2. ¿Hubo un incremento en la resistencia de las baldosas al cambiar de proveedor, tomando en cuenta un nivel de significación de 0.02?

27. El área de riesgos de una aseguradora sabe que en promedio los conductores necesitan algún servicio de la empresa cada 2 meses; sin embargo en una muestra aleatoria de 45 clientes se observa que utilizan por lo menos su seguro cada 1.8 meses con una desviación estándar de 0.5. Tomando en cuenta un nivel de significación de 0.05, determine si el promedio es el mismo que indica el área de riesgos de la empresa.

28. Una empresa dedicada al envío de mensajería a nivel metropolitano sabe que el tiempo promedio de entrega de sus paquetes es de 20 min. Desde hace 2 meses se cambió el equipo de transporte y se espera que el tiempo de entrega sea menor, por lo que se toma una muestra aleatoria de 77 entregas y se observa que en promedio el tiempo es de 17.5 min con una desviación estándar de 2.5 min. Determine, de acuerdo con un nivel de significación de 0.06, si disminuyó el tiempo promedio de entrega.

29. Los datos del mercado bursátil indican que todos los instrumentos registrados aumentaron en promedio $1.2, sin embargo un analista bursátil dice que ese dato es errado y que el verdadero es superior. Una muestra aleatoria de 50 instrumentos demuestra que en promedio subieron $1.3 con una desviación estándar de $0.30. ¿Los datos del mercado son verídicos o el analista tiene razón? Considere un nivel de significación de 0.01.

9.8 Prueba de hipótesis sobre una proporción poblacional de 2 colas

30. Una distribuidora entrega en tiempo y forma en promedio 70% de sus pedidos; para una evaluación interna se toma una muestra aleatoria de 200 pedidos y se encuentra que 150 son entregados correctamente. Determine si los datos de la muestra corresponden al porcentaje promedio de entrega de pedidos con un nivel de significación de 0.05.

31. Una embotelladora tiene como regla que 25% de sus envases deben ser elaborados con material reciclado, por lo que el jefe de producción toma una muestra aleatoria de 85 envases para comprobar que se esté cumpliendo con ese requisito, y encuentra que 23.5% están hechos con *pet* reciclado. Tomando en cuenta un nivel de significación de 0.03, determine si la regla se está cumpliendo en esta empresa.

32. A un grupo de deportistas de alto rendimiento se le exige que durante el entrenamiento consuma bebidas energéticas con sólo 6% de sodio; los entrenadores toman una muestra aleatoria de 36 deportistas y encuentran que las bebidas que toman son de 600 ml donde 84 ml son de sodio. ¿Están cumpliendo los deportistas con lo exigido, de acuerdo con un nivel de significación de 0.07?

33. Un maestro está preparando a un grupo de alumnos para un maratón de matemáticas. De acuerdo con el plan de estudios, los alumnos deben estar en el taller 20% de su horario escolar; una muestra aleatoria de 55 alumnos determina que de 8 h de clase los alumnos acuden al taller 1 h. Identifique si los alumnos están cumpliendo con el tiempo de estudio para la preparación del maratón con un nivel de significación de 0.01.

34. El gerente de recursos humanos de una empresa evalúa cada mes el rendimiento de los empleados de acuerdo con el cumplimiento de objetivos; para obtener una calificación satisfactoria es necesario cumplir 85% de las metas fijadas a principio de cada mes. Una muestra aleatoria de 120 empleados refleja que de 15 objetivos, se cumplieron 11. Determine si la calificación de los empleados es satisfactoria de acuerdo con un nivel de significación de 0.04.

Cola inferior, lado izquierdo

35. En una empresa se imprimen diariamente 2 500 hojas de las cuales 750 son circulares internas. Con la intención de disminuir esta proporción se exige a los gerentes mandar la mayor parte de sus circulares a través de correo electrónico. Una muestra aleatoria de 67 días demuestra que las circulares publicadas corresponden a 15% del total de impresiones diarias. Determine, con un nivel de significación de 0.02, si hubo efectividad en la intención de disminuir la proporción de circulares impresas diariamente.

36. El departamento de recursos humanos de una empresa registra que 8% de aspirantes se incorporan anualmente a trabajar en la compañía. Este año la oferta de trabajo es menor, de una muestra aleatoria de 350 personas que aplicaron para obtener una vacante, 24 se quedaron en la empresa. Tomando en cuenta un nivel de significación de 0.08 determine si se respetó el ajuste en la disminución de plazas de trabajo.

37. Una fábrica de ropa compra normalmente tela compuesta por 20% de poliéster y 80% de algodón para producir diferentes prendas; el próximo mes saldrá a la venta ropa para niños y es muy importante disminuir el porcentaje de poliéster en la tela, por lo que cambian de proveedor. Una muestra aleatoria de 300 m de tela muestra que su composición es 88% algodón y el resto poliéster. ¿Hubo una disminución significativa en la cantidad de poliéster en la tela con un nivel de significación de 0.01?

38. El director académico de una escuela sabe que de cada 10 alumnos 4 reprueban matemáticas, por lo que implementa cursos extracurriculares programados por las tardes; al finalizar el año, en una muestra aleatoria de 250 niños se observa que 75 reprobaron la materia. Con un nivel de significación de 0.09, determine si disminuyó el porcentaje de niños que reprueba matemáticas.

39. Una delegación tiene como dato que 27% del total de la población con capacidad para votar no lo hace, por lo que desarrolla una campaña para incentivar a las personas a votar por sus representantes. Una muestra aleatoria de 800 personas demuestra que 110 con posibilidad de votar no lo hicieron. Determine, con un nivel de significación de 0.03, si la campaña cumplió su objetivo.

Cola superior, lado derecho

40. Una universidad tiene registrado que 5% de sus alumnos foráneos viven en las residencias del campus. Como existe espacio en estas residencias para alumnos adicionales y la vida allí les ofrece ventajas, se implanta un programa de promoción de esos espacios y al siguiente semestre se toma una muestra aleatoria de 80 estudiantes foráneos y se encuentra que 9 de ellos viven en las residencias universitarias. Determine, con un nivel de significación de 0.04, si se incrementó el uso de este espacio.

41. El gobierno de una ciudad sabe que de cada 10 familias 3 toman vacaciones en semana santa y para este año hizo gran publicidad de los destinos turísticos más representativos. Una muestra aleatoria de 600 familias refleja que 32% de ellas tuvieron un plan vacacional fuera de la ciudad. Con un nivel de significación de 0.05, determine si la publicidad cumplió su función.

42. Una institución financiera sabe que sólo 50% de los clientes que gozan de un préstamo pagan la cuota correspondiente en la fecha indicada, por lo que se decidió implementar mejorías (capacitación y cambio de supervisor) al área de cuentas por cobrar. Una muestra aleatoria de 180 clientes deudores señala que 105 de ellos pagan en tiempo y forma lo que les corresponde. Determine, con un nivel de significación de 0.07, si las mejorías en el área hicieron efecto.

43. El gerente de un club deportivo sabe que sólo 32% de los socios acuden al restaurante del lugar, así que durante un mes se propone hacer que todos los socios conozcan la comida y el servicio del restaurante. En una muestra aleatoria de 100 socios se observa que 38 de ellos van al restaurante, ¿hubo incremento en la proporción de socios que usan el restaurante del club? Considere un nivel de significación de 0.02.

44. Un estudio agrícola demuestra que 75% de la planta X crece en menos de 2 meses. Se comenzó a utilizar un fertilizante con la intención de incrementar esta proporción. Una muestra aleatoria de 800 plantas señala que 420 crecen en menos de 2 meses. Determine, con un nivel de significación de 0.01, si el porcentaje se incrementó.

capítulo 10

Pruebas de hipótesis para 2 poblaciones

Sumario

10.1 Panorama general de las pruebas de hipótesis
10.2 Pruebas de hipótesis sobre la diferencia entre 2 medias
10.2.1 Pruebas con muestras grandes e independientes
10.2.2 Pruebas con muestras pequeñas e independientes, variables distribuidas normalmente
10.2.3 Pruebas para muestras pareadas cuando no se conocen las varianzas pero no se necesita asumir que sean iguales
10.3 Pruebas de hipótesis sobre la diferencia entre 2 proporciones
10.4 Prueba para la diferencia entre 2 varianzas
10.4.1 Distribución F y Excel
10.5 Excel y pruebas de hipótesis para 2 muestras
10.6 Resumen
10.7 Fórmulas del capítulo
10.8 Ejercicios adicionales

En el capítulo anterior se revisaron los procedimientos que se utilizan para realizar pruebas de hipótesis sobre un parámetro poblacional: una media y una proporción. En éste se revisarán los procedimientos para probar hipótesis sobre los parámetros de 2 poblaciones. Asimismo se estudiarán los métodos para realizar pruebas —de diversas variaciones— sobre la diferencia entre 2 medias, 2 proporciones y 2 varianzas poblacionales.

Resulta conveniente comenzar este capítulo con un panorama general de las pruebas de hipótesis, de las que existen diversas variaciones. Esta introducción incluye tanto a casos del presente capítulo como del anterior; además, contempla una panorámica de todas las pruebas de hipótesis que se analizan en el libro. Es oportuno mencionar que en otros capítulos se estudiarán más tipos de pruebas de hipótesis.

Así, en la sección siguiente se presenta este conglomerado general de las pruebas de hipótesis, mientras que en el tema 10.2 se revisarán los detalles de cada una de ellas para 2 medias. En las secciones 10.3 y 10.4 se analizarán los procedimientos para pruebas de 2 proporciones y de 2 varianzas, en ese orden.

Al igual que en los demás capítulos, se termina con secciones sobre uso de Excel, un resumen y una lista de las fórmulas que se introducen en el capítulo.

10.1 Panorama general de las pruebas de hipótesis

Vale la pena insistir en que posiblemente los métodos de pruebas de hipótesis son una de las metodologías estadísticas más importantes y útiles. Además, es un tema considerablemente amplio y, por ello, conviene hacer una pausa para presentar un panorama general que proporcione una idea global sobre el tema para visualizar mejor el terreno.

En el capítulo anterior se presentaron las pruebas de hipótesis para una muestra y se incluyó una introducción. Conviene presentar esquemáticamente ese contenido:

Conceptos básicos:

- Hipótesis nula H_0, e hipótesis alternativa H_1.
- Errores Tipo I: rechazar una hipótesis cierta.
  Tipo II: aceptar una hipótesis falsa.

Pruebas de hipótesis para una muestra:

- Sobre una media, μ, para muestras grandes y pequeñas.
- Sobre una proporción, π, para muestras grandes.
- Pueden ser de 1 o de 2 extremos, lo cual define la región de aceptación y la o las regiones de rechazo.

Tres métodos para realizar pruebas de hipótesis:

1. Método del intervalo (que se asocia directamente con la técnica de estimación de parámetros).
2. Método del estadístico de prueba (se vieron la z de la normal y la t de Student).
3. Método de P: la probabilidad.

Los tres métodos conducen a los mismos resultados utilizando de forma distinta los mismos elementos.

El procedimiento para realizar pruebas de hipótesis consiste en:

1. Plantear H_0 y H_1.
2. Determinar α, el nivel de significación.
3. Elaborar una gráfica (generalmente muy útil).
4. Identificar las regiones de aceptación y de rechazo.
5. Calcular el error estándar del estadístico.
6. Tomar la decisión.
7. Interpretar los resultados.

En este capítulo se presentarán las técnicas para realizar pruebas de hipótesis para 2 poblaciones; en seguida se enlistan los casos que se analizarán con sus respectivas particularidades:

1. Pruebas de hipótesis sobre la diferencia entre 2 medias.

   a) Pruebas con muestras grandes e independientes.

   - Cuando se conocen las varianzas de las 2 poblaciones.
   - Cuando no se conocen las varianzas y no se asume que sean iguales.
   - Cuando no se conocen las varianzas pero se asume que son iguales.

   b) Pruebas con muestras pequeñas e independientes.

   - Cuando no se conocen las varianzas pero se asume que son iguales.
   - Cuando no se conocen las varianzas y no se asume que sean iguales.

   c) Pruebas para muestras pareadas cuando no se conocen las varianzas pero no se necesita asumir que sean iguales.

2. Pruebas de hipótesis sobre la diferencia entre 2 proporciones.
3. Pruebas de hipótesis sobre la diferencia entre 2 varianzas.

En el siguiente capítulo, "Pruebas de hipótesis con la distribución χ^2 (ji cuadrada)", se comienza por analizar esta distribución ji cuadrada, otra distribución teórica de probabilidad como la normal y la t de Student. Se analizarán las características de esta distribución y se continúará con su aplicación para realizar las siguientes pruebas de hipótesis:

- Pruebas de hipótesis para la varianza de una población.
- Pruebas para la diferencia entre 2 proporciones (prueba de homogeneidad).
- Prueba para la diferencia entre n proporciones (prueba de homogeneidad).
- Pruebas de bondad de ajuste de una distribución empírica a:
  - una distribución normal.
  - una distribución binomial.
  - una distribución Poisson.
- Pruebas sobre la independencia entre dos variables.

En el capítulo 12 se abordará el tema de *Análisis de varianza* que se ocupa de analizar diversas técnicas para probar hipótesis sobre la posible igualdad entre más de 2 medias.

En los capítulos 13 y 14 que tocarán el análisis de regresión simple y múltiple, respectivamente, se incluirán secciones que aborden pruebas de hipótesis principalmente sobre la pendiente, o coeficiente β, de una ecuación de regresión lineal. Por supuesto no se pretende que se entienda ahora de qué se trata esto

pero es importante mencionar que también en este tema se utilizan pruebas de hipótesis y, en todo caso, en esos capítulos se explicarán los detalles.

El capítulo 17, "La estadística no paramétrica", es un apartado enteramente dedicado a pruebas de hipótesis que tienen características muy diferentes a las explicadas previamente (a excepción de las pruebas con X^2 del capítulo 11). Algunas de estas pruebas de hipótesis pueden ser clasificadas como pruebas no paramétricas:

- Pruebas de bondad de ajuste de una distribución empírica a:
  - Una distribución normal.
  - Una distribución binomial.
  - Una distribución Poisson.
- Pruebas sobre la independencia entre 2 variables:

En este capítulo se estudian pruebas de hipótesis para aleatoriedad y las pruebas sobre medianas (nótese que no son medias, sino medianas) que pueden aplicarse cuando los datos disponibles no son numéricos, sino que están en escala categórica ordinal.

Como puede verse en este panorama general, el uso de pruebas de hipótesis tiene abundantes aplicaciones. Vale la pena tener esto presente al estudiar cualquiera de los temas que se aborden, ya que mejorará la comprensión de las técnicas de pruebas de hipótesis debido a que todas comparten muchos rasgos comunes.

10.2 Pruebas de hipótesis sobre la diferencia entre 2 medias

En ocasiones, lo que se desea probar es si existen diferencias entre los parámetros de 2 poblaciones. Por ejemplo, puede desearse probar si existe diferencia entre el promedio de artículos producidos en el turno matutino (población 1) y el turno vespertino (población 2). También, puede ser necesario probar si la proporción de hombres en una ciudad (población 1) es igual a la proporción de hombres en otra ciudad (población 2).

Aunque apenas son 2 ejemplos, se puede sospechar que la lista de posibles casos es enorme. En el primer ejemplo se utilizaría una prueba para la diferencia entre 2 medias, en tanto que en el segundo caso se emplearía una prueba para la diferencia entre 2 proporciones.

Tal como se verá en las secciones siguientes, el capítulo está dividido precisamente en esos 2 tipos de pruebas, y además se incluye una sección adicional sobre pruebas para 2 varianzas, misma que servirá para conectar este capítulo con el siguiente. En cada tipo de pruebas se presentan distintas circunstancias entre las que sobresalen las aplicaciones con muestras grandes o con muestras pequeñas. Del mismo modo que sucedió en las pruebas de hipótesis para una muestra en el capítulo anterior, y de la misma manera que sucedió en el tema de estimaciones de parámetros, tratar con muestras grandes ($n \geq 30$) o muestras pequeñas ($n < 30$) implica diferencias en los procedimientos.

A su vez en las pruebas de hipótesis sobre medias, además del asunto del tamaño de la muestra, es necesario considerar si se conocen o no las varianzas poblacionales y si puede o no asumir que son iguales.

Asimismo se abordan las consideraciones necesarias a tomar en cuenta cuando se trata de muestras independientes (la mayoría de los casos aquí presentados) o de **datos pareados**, es decir, dependientes, ya que esto también implica cambios en los procedimientos.

Datos pareados. Datos dependientes.

En las secciones siguientes se revisarán todos estos casos, ahora se muestran los pares de hipótesis que se manejan para la diferencia entre 2 medias:

Para una prueba de 2 extremos:

$$H_0: \mu_1 = \mu_2$$
$$H_1: \mu_1 \neq \mu_2$$

O, de manera equivalente:

$$H_0: \mu_1 - \mu_2 = 0$$
$$H_1: \mu_1 - \mu_2 \neq 0$$

Para pruebas de un extremo:

$$H_0: \mu_1 \geq \mu_2$$
$$H_1: \mu_1 < \mu_2$$

Equivalencia:

$$H_0: \mu_1 - \mu_2 \leq 0$$
$$H_1: \mu_1 - \mu_2 > 0$$

Nótese que pueden plantearse las hipótesis como:

- Una igualdad o desigualdad entre 2 medias, o de manera alternativa y equivalente.
- Una diferencia entre medias que es igual, mayor o menor que cero.

En las secciones siguientes se revisarán los distintos casos.

10.2.1 Pruebas con muestras grandes e independientes

En esta sección se presentan casos de pruebas de hipótesis con 2 características: las muestras son grandes, tanto n_1 como n_2 son mayores que o iguales a 30, y se trata de muestras independientes, es decir, muestras tomadas de 2 poblaciones diferentes.

En las subsecciones siguientes se definen las hipótesis, los estadísticos de prueba y los errores estándar porque, como se recordará, el procedimiento para realizar pruebas de hipótesis es:

1. Plantear H_0 y H_1.
2. Determinar α, el nivel de significación.
3. Elaborar una gráfica (por lo general muy útil).
4. Identificar las regiones de aceptación y de rechazo.
5. Calcular el error estándar del estadístico.
6. Tomar la decisión.
7. Interpretar los resultados.

Los diferentes casos de pruebas para 2 medias implican diferencias en el estadístico de prueba pero, sobre todo, en la forma en la que se calcula el error estándar del estadístico, es decir, la media de la distribución muestral del estadístico.

10.2.1.1 Cuando se conocen las varianzas de las 2 poblaciones

Si se trata de muestras grandes e independientes y si se conocen las verdaderas varianzas de las poblaciones correspondientes, el estadístico de prueba es la ya conocida z estandarizada de la distribución normal que para 2 poblaciones se calcula como:

$$z = \frac{\left(\bar{X}_1 - \bar{X}_2\right) - \left(\mu_1 - \mu_2\right)}{\sigma_{X_1 - X_2}}$$

Pero como la hipótesis nula plantea que:

$$H_0: \mu_1 - \mu_2 = 0$$

La expresión anterior se convierte en:

$$z = \frac{\bar{X}_1 - \bar{X}_2}{\sigma_{X_1 - X_2}} \tag{10.1}$$

En donde el error estándar de la diferencia entre 2 medias es:

$$\sigma_{X_1 - X_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \tag{10.2}$$

Sin embargo, el caso más común es que no se conozcan las varianzas, entonces se utilizan las de las muestras para estimarlas, y el procedimiento es exactamente igual. Las fórmulas que describen al estadístico de prueba y al error estándar de la diferencia entre las 2 medias cambian simplemente como se describe en el apartado siguiente.

10.2.1.2 Cuando no se conocen las varianzas y no se asume que sean iguales

Como se menciona en el párrafo anterior, la única diferencia entre las fórmulas para calcular el estadístico de prueba y el error estándar de la diferencia entre 2 medias, cuando se utilizan datos muestrales es que se sustituye s^2 por σ^2 y $s_{X_1 - X_2}$ por $\sigma_{X_1 - X_2}$, de la siguiente manera:

$$z = \frac{\bar{X}_1 - \bar{X}_2}{s_{X_1 - X_2}} \tag{10.3}$$

En donde:

$$s_{X_1 - X_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{10.4}$$

EJEMPLO 10.1

Un gerente de producción desea determinar si existe diferencia entre la productividad de los trabajadores del turno matutino y los del turno vespertino. Para ello, toma una muestra aleatoria de 30 trabajadores de cada turno y encuentra que produjeron un promedio de 68 artículos por turno, con una desviación estándar de 16, en tanto que los del turno vespertino produjeron 65.5 artículos en promedio con desviación estándar de 17. ¿Existe diferencia entre la productividad de los 2 turnos, a un nivel de significación de 0.01?

Solución: En primer lugar, las hipótesis:

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

Como $\alpha = 0.01$ y se trata de una prueba de 2 extremos de la tabla de áreas bajo la curva normal, el valor crítico del estadístico de prueba es ±2.575, ya que:

$$P(-2.575 \leq z \leq 2.575) = 0.1$$

Se calcula ahora el error estándar de la diferencia entre 2 medias con la fórmula anotada antes:

$$s_{X_1 - X_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{16^2}{30} + \frac{17^2}{30}} = \sqrt{8.53 + 9.63} = 4.26$$

El valor calculado del estadístico de prueba es:

$$z = \frac{\bar{X}_1 - \bar{X}_2}{s_{X_1 - X_2}} = \frac{68 - 65.5}{4.26} = 0.59$$

Este valor observado del estadístico de prueba está dentro del rango crítico de –2.575 a 2.575, así que no se tienen elementos para rechazar la hipótesis nula, por lo que se concluye que la producción promedio en los 2 turnos es igual.

10.2.1.3 Cuando no se conocen las varianzas pero se asume que son iguales

En estas condiciones (recuérdese que también se trata de muestras grandes e independientes), el estadístico de prueba sigue siendo la z de la distribución normal estándar:

$$z = \frac{\bar{X}_1 - \bar{X}_2}{s_{X_1 - X_2}}$$

Pero ahora, como se supone que las varianzas de las 2 poblaciones son iguales, se combinan las varianzas muestrales de la siguiente manera:

$$S_c^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \tag{10.5}$$

Nótese que esta forma de combinar las varianzas muestrales es una forma de ponderación, en donde los pesos son los respectivos tamaños de muestra. Una vez realizada la combinación de varianzas, se calcula el error estándar de la diferencia de medias de la misma manera que se hizo antes pero ahora utilizando la varianza combinada, s_c^2:

$$s_{X_1 - X_2} = \sqrt{\frac{s_c^2}{n_1} + \frac{s_c^2}{n_2}} \tag{10.6}$$

EJEMPLO 10.2

Se desea probar si el salario medio mensual de los empleados oficinistas de 2 empresas del ramo de servicios turísticos son iguales o no, con un nivel de significación de 1%. Para ello, se toman muestras de ambas empresas y los datos correspondientes se resumen en el siguiente cuadro:

	Muestra de la empresa 1	Muestra de la empresa 2
Tamaño, n	$n_1 = 50$	$n_2 = 60$
Media	$\bar{X}_1 = 6\,000$	$\bar{X}_2 = 5\,850$
Desviación estándar	$S_1 = 300$	$S_2 = 214$

Solución: Las hipótesis:

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

O de manera equivalente:

$$H_0 : \mu_1 - \mu_2 = 0$$
$$H_1 : \mu_1 - \mu_2 \neq 0$$

Aquí $\alpha = 0.01$ y, como se trata de una prueba de 2 extremos, el valor crítico del estadístico de prueba z es:

$$P(-2.575 \leq z \leq 2.575) - 0.99$$

O igualmente:

$$P(-2.575 \geq z \geq 2.575\ ) = 0.01$$

El valor de z calculado con los datos muestrales es:

$$z = \frac{\bar{X}_1 - \bar{X}_2}{s_{X_1 - X_2}}$$

En donde:

$$S_c^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{(50 - 1)(300)^2 + (60 - 1)(214)^2}{50 + 60 - 2}$$
$$= \frac{4\,410\,000 + 2\,701\,964}{108} = 65\,851.52$$

Y:

$$s_{X_1 - X_2} - \sqrt{\frac{s_c^2}{n_1} + \frac{s_c^2}{n_2}} - \sqrt{\frac{65\,851.52}{50} + \frac{65\,851.52}{60}}$$
$$- \sqrt{2\,414.56} - 49.14$$

Por lo que:

$$z = \frac{\bar{X}_1 - \bar{X}_2}{s_{X_1 - X_2}} = \frac{6\,000 - 5\,850}{49.14} = \frac{150}{49.14} = 3.05$$

Como el valor calculado de z cae en la región de rechazo, es decir, es mayor que la z crítica de 2.575, se rechaza la hipótesis nula y se concluye que los salarios promedio mensuales de los empleados oficinistas de las 2 empresas de servicios turísticos no son iguales.

10.2.2 Pruebas con muestras pequeñas e independientes, variables distribuidas normalmente

Las circunstancias son, en este caso, diversas:

- La variable se distribuye de forma normal en ambas poblaciones.
- Las 2 muestras son independientes.
- Los tamaños de las muestras son pequeños: $n_1 < 30$ y $n_2 < 30$.
- No se conocen las varianzas de las poblaciones correspondientes.

En seguida se revisarán los procedimientos de pruebas de hipótesis para estas circunstancias y para 2 casos distintos: cuando puede asumirse que las varianzas poblacionales son iguales y cuando debe aceptarse que son distintas.

10.2.2.1 Cuando no se conocen las varianzas pero se asume que son iguales

En estas circunstancias, el estadístico de prueba apropiado es la t de Student:

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)0}{s_{X_1 - X_2}}$$

Pero, de nuevo, como la hipótesis nula plantea que:

$$H_0 : \mu_1 - \mu_2 = 0$$

La expresión anterior se convierte en:

$$t = \frac{(\bar{X}_1 - \bar{X}_2)}{s_{X_1 - X_2}} \tag{10.7}$$

Con $n_1 + n_2 - 2$ grados de libertad.

Al igual que antes, cuando es asumido que las 2 varianzas poblacionales son iguales, éstas se combinan, como en la fórmula anterior (10.5):

$$s_c^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

El error estándar de la diferencia entre 2 medias, que es la fórmula anterior (10.6):

$$s_{X_1 - X_2} = \sqrt{\frac{s_c^2}{n_1} + \frac{s_c^2}{n_2}}$$

Ejemplo 10.3

Un departamento de control de calidad desea evaluar 2 máquinas que fabrican ciertas piezas circulares de plástico. Se desea saber si la máquina A las fabrica con un diámetro mayor que la máquina B. Para ello, se toma una muestra de 12 piezas de la máquina A y se encuentra que su diámetro tiene una media de 1.061 cm con varianza de 0.000442. En tanto que una muestra aleatoria de 10 piezas de la máquina B arroja una media de 1.038 cm con varianza de 0.000228. Si los diámetros de estas piezas se distribuyen de forma normal en las 2 máquinas y se sabe que sus varianzas son iguales, compruebe la hipótesis de que la máquina A está fabricando piezas de mayor diámetro, con un nivel de significación de 0.05.

Solución: Las hipótesis:

$$H_0: \mu_A \leq \mu_B$$
$$H_1: \mu_A > \mu_B$$

Aquí $\alpha = 0.05$ y como se trata de una prueba de un extremo donde se tienen $n_1 + n_2 - 2 = 12 + 10 - 2 = 20$ grados de libertad, el valor crítico del estadístico de prueba t es:

$$P((\, t \geq 1.7247 \mid gl = 20\,)) = 0.05$$

Por su parte, el valor de t calculado con los datos muestrales es:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{X_1 - X_2}}$$

En donde:

$$s_c^2 = \frac{(n_1 - 1)^{s_1^2} + (n_1 - 1)^{s_2^2}}{n_1 + n_2 - 2}$$

$$= \frac{(12 - 1)(0.000442) + (10 - 1)(0.000228)}{12 + 10 - 2}$$

$$= \frac{0.004862 + 0.002052}{20} = 0.0003457$$

Y,

$$s_{X_1 - X_2} = \sqrt{\frac{s_c^2}{n_1} + \frac{s_c^2}{n_2}} = \sqrt{\frac{0.0003457}{12} + \frac{0.0003457}{10}}$$

$$= \sqrt{0.000063378} = 0.00796$$

Por lo que:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{X_1 - X_2}} = \frac{1.061 - 1.038}{0.00796} = \frac{0.023}{0.00796} = 2.89$$

Así que como el valor calculado de t, 2.89, es mayor que el valor crítico determinado según el nivel de significación, 1.7247, se rechaza la hipótesis nula para concluir que, efectivamente, la máquina A está fabricando piezas con mayor diámetro que las producidas por la máquina B.

10.2.2.2 Cuando no se conocen las varianzas y no se asume que sean iguales

En este caso, el estadístico de prueba sigue siendo la t de Student pero se trata de una t modificada. Se ponderan los valores de t correspondientes a cada muestra mediante el cociente entre sus correspondientes varianza y tamaño de muestra, como se aprecia en la siguiente fórmula de t crítica ponderada:

$$t'_{cr} = \frac{\frac{s_1^2}{n_1} t_1 + \frac{s_2^2}{n_2} t_2}{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \qquad \textbf{(10.8)}$$

Por su parte, al igual que se hizo antes, la t calculada es:

$$t = \frac{\left(\bar{X}_1 - \bar{X}_2\right) - \left(\mu_1 - \mu_2\right)0}{s_{X_1 - X_2}}$$

Pero, de nuevo, como la hipótesis nula plantea que:

$$H_0: \mu_1 - \mu_2 = 0$$

La expresión anterior corresponde a la fórmula 10.7:

$$t = \frac{\left(\bar{X}_1 - \bar{X}_2\right)}{s_{X_1 - X_2}}$$

Con $n_1 + n_2 - 2$ grados de libertad.

Además, como no se asume que las varianzas sean iguales, no se combinan las 2 varianzas muestrales y, entonces, el error estándar se calcula simplemente como:

$$s_{X_1 - X_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Que es la misma fórmula 10.4 anterior.

Ejemplo 10.4

De acuerdo con el ejemplo anterior, suponga que no es posible asumir que las varianzas de las 2 poblaciones sean iguales. En este caso, la prueba se realiza tal como se ilustra en seguida:

Solución: Los datos:

	Máquina A	Máquina B
n	12	10
$\bar{X}$	1.061	1.038
s^2	0.000442	0.000228

Las hipótesis:

$$H_0: \mu_A \leq \mu_B$$
$$H_1: \mu_A > \mu_B$$

Aquí, con $\alpha = 0.05$ y, dadas las circunstancias, el valor crítico del estadístico de prueba se calcula de la siguiente manera (nótese que los grados de libertad son el tamaño de la muestra, n, menos 1):

$$P(t_1 > 1.7959 \mid gl = 11) = 0.05$$
$$P(t_2 > 1.8331 \mid gl - 9) - 0.05$$

Y,

$$t'_{cr} = \frac{\frac{s_1^2}{n_1}t_1 + \frac{s_2^2}{n_2}t_2}{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \frac{1.7959\frac{0.000442}{12} + 1.8331\frac{0.000228}{10}}{\frac{0.000442}{12} + \frac{0.000228}{10}}$$

$$= \frac{0.000066148 + 0.000041794}{0.000036833 + 0.0000228} = \frac{0.000107942}{0.000059633} = 1.81$$

Ahora, el error estándar de la diferencia entre 2 medias, en estas circunstancias:

$$s_{X_1 - X_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{0.000442}{12} + \frac{0.000228}{10}}$$

$$= \sqrt{0.00005963} = 0.007722$$

Finalmente, la t calculada:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{X_1 - X_2}} = \frac{1.061 - 1.038}{0.007722} = \frac{0.023}{0.007722} = 2.98$$

Ya que esta t calculada es mayor que la t crítica, cae en la región de rechazo; se rechaza la hipótesis nula y se concluye que, efectivamente, la máquina A está fabricando piezas con mayor diámetro que la máquina B.

10.2.3 Pruebas para muestras pareadas cuando no se conocen las varianzas pero no se necesita asumir que sean iguales

Se analizó el caso de una prueba para la diferencia entre 2 medias provenientes de poblaciones independientes. Aquí se analizará el caso de la diferencia entre 2 medias provenientes de poblaciones pareadas o relacionadas. Es importante tener presentes las circunstancias de estos casos:

- Se trata de muestras pareadas.
- Los tamaños de muestras son pequeños.
- La variable se distribuye de forma normal en la población.

En este caso, la prueba se convierte en una prueba sobre la diferencia entre las observaciones, ya que se calculan las diferencias entre:

1. Dos individuos de la misma especie sometidos a tratamientos diferentes (pareamiento de individuos según una característica de interés).
2. Dos mediciones hechas a los mismos individuos.

La media de las diferencias es:

$$\bar{D} = \frac{\sum D_i}{n} \tag{10.9}$$

Con el teorema central del límite, el promedio de las diferencias sigue una distribución normal cuando se conoce la varianza de las diferencias y n es grande. Pero generalmente no se conoce la varianza de las diferencias, entonces se le estima:

$$S = \sqrt{\frac{\sum_{i=1}^{n}\left(D_i - \bar{D}\right)^2}{n-1}} \tag{10.10}$$

El error estándar de las diferencias pareadas es:

$$s_D = \frac{s}{\sqrt{n}} \tag{10.11}$$

Con muestras pequeñas, el estadístico de prueba es:

$$t_{n-1} = \frac{\bar{D}}{s_D} \tag{10.12}$$

Con $n - 1$ grados de libertad. Nótese que aquí cambian los grados de libertad, al tratarse de muestras pareadas.

Ejemplo 10.5

Un fabricante de neumáticos para automóvil desea evaluar si es significativa la diferencia entre la duración de cierto tipo de llanta que corre a diferentes velocidades, toma 2 muestras de 10 neumáticos de ese tipo y las prueba a velocidades de 80 y 100 km/h. Los resultados se muestran a continuación, se trata de miles de km recorridos:

Par de llantas	80 km/h	100 km/h
1	42.27	38.90
2	54.03	50.03
3	56.67	49.14
4	48.51	45.82
5	36.56	37.76
6	68.34	58.26
7	40.14	34.34
8	50.82	48.99
9	47.84	45.28
10	45.60	44.64

Se desea probar si la duración de los neumáticos es mayor a menores velocidades, con un nivel de significación de 0.01.

Solución: Las hipótesis:

$$H_0: \mu_D \geq 0 \text{ (es decir } \mu_{100} \geq \mu_{80}\text{)}$$
$$H_1: \mu_D < 0 \text{ (es decir } \mu_{100} < \mu_{80}\text{)}$$

Si el nivel de significación, $\alpha = 0.01$, el valor crítico del estadístico de prueba es:

$$P(t < -2.821 \mid gl = 9) = 0.01$$

Las operaciones para calcular la media de las diferencias se resumen en el siguiente cuadro:

$$\bar{D} = \frac{\sum D_i}{n} = \frac{-27.62}{10} = -2.762$$

100 km/h	80 km/h	D_i	$D_i - \bar{D}$	$(D_i - \bar{D})^2$
38.9	42.27	–3.37	–0.61	0.37
50.03	54.03	–4.00	–1.24	1.53
49.14	56.67	–7.53	–4.77	22.73
45.82	48.51	–2.69	0.07	0.01
37.76	36.56	1.20	3.96	15.70
58.26	58.34	–0.08	2.68	7.19
34.34	40.14	–5.80	–3.04	9.23
48.99	50.82	–1.83	0.93	0.87
45.28	47.84	–2.56	0.20	0.04
44.64	45.6	–0.96	1.80	3.25
		–27.62		60.92

$$s = \sqrt{\frac{\sum_{i=1}^{n}\left(D_i - \bar{D}\right)^2}{n-1}} = \sqrt{\frac{60.92}{10-1}} = \sqrt{6.769} = 2.602$$

El error estándar de las diferencias:

$$s_D = \frac{s}{\sqrt{n}} = \frac{2.602}{\sqrt{10}} = \frac{2.602}{3.16} = 0.823$$

Ahora, el valor calculado del estadístico de prueba:

$$t_{n-1} = \frac{\bar{D}}{s_D} = \frac{-2.762}{0.823} = -3.36$$

Entonces, como el valor calculado de t es menor que su valor crítico, se rechaza la hipótesis nula y se concluye que los neumáticos usados a 80 km/h duran más que los que corren a 100 km/h.

EJERCICIOS 10.2 Pruebas de hipótesis sobre la diferencia entre 2 medias

10.2.1 Pruebas con muestras grandes e independientes

10.2.1.2 Cuando no se conocen las varianzas y no se asume que sean iguales

1. En 2 ciudades en las que existen refinerías se tomó una muestra a cada persona de un grupo de 35, y se midió el nivel de plomo en la sangre. En la ciudad A se encontró que en promedio el nivel de plomo es de 79.4 microgramos con una desviación estándar de 8. En la ciudad B, el promedio es de 78 microgramos con una desviación estándar de 1. ¿Existe diferencia en el nivel de plomo en la sangre de los habitantes de cada ciudad, a un nivel de significación de 0.01?
2. Para la fabricación de una pieza específica se emplean 2 máquinas. Se toma una muestra de 40 piezas elaboradas por ambos aparatos y se encuentra que las piezas que produjo la máquina A tienen una longitud promedio de 83 mm con una desviación estándar de 5 mm, mientras que las de la máquina B la longitud promedio es de 82 mm con una desviación estándar de 2 mm. Determine si existe diferencia entre la longitud de las piezas fabricadas por cada máquina, con una nivel de significación de 5 por ciento.
3. En una facultad se imparten 2 licenciaturas, administración y contaduría. Se preguntó a 42 alumnos de administración cuál es el número de veces que han consultado libros en la biblioteca durante el último mes, y se obtuvo que en promedio fueron 27 veces con una desviación estándar de 4, mientras que en la muestra de 37 alumnos de contaduría el promedio fue de 23 con una desviación estándar de 3. Compruebe si existe diferencia entre el promedio de consultas realizadas por los alumnos de cada licenciatura con un nivel de significación de 0.05.
4. Para probar la velocidad de combustión de 2 tipos de aceite automotriz se tomó una muestra de 50 botellas de 2 marcas distintas. En la marca A se encontró que el tiempo promedio de combustión es de 47.5 seg con una desviación estándar de 3.2 seg, y en la marca B el tiempo promedio es de 49.4 seg con una desviación estándar de 3.7 seg. Compruebe si existe diferencia entre el tiempo de combustión de las 2 marcas de aceite con un nivel de significación de 0.01.
5. Una compañía telefónica brinda 2 tipos de servicios, plan y prepago, y desea saber si existe diferencia entre el número de minutos utilizados mensualmente en cada servicio. En el caso de los usuarios del servicio de plan se tomó una muestra de 36 personas y se encontró que el promedio de minutos fue de 237 con una desviación estándar de 8.7. De los usuarios de prepago se tomó una muestra de 41 y el promedio fue de 248 con una desviación estándar de 10.4. Compruebe la hipótesis con un nivel de significación del 0.01.

10.2.1.3 Cuando no se conocen las varianzas pero se asume que son iguales

6. Se desea probar si el peso promedio de los niños que cursan el tercer grado de primaria en 2 escuelas diferentes es igual o no, con un nivel de significación de 1%. Para ello se toman muestras en ambas escuelas y se obtienen los siguientes datos:

	Muestra de la escuela 1	Muestra de la escuela 2
Tamaño, n	$n_1 = 30$	$n_2 = 33$
Media	$\overline{X}_1 = 33$	$\overline{X}_2 = 31.5$
Desviación estándar	$S_1 = 2.8$	$S_1 = 3.4$

7. Se está experimentando con un nuevo fertilizante orgánico para la siembra de trigo con la finalidad de probar si el crecimiento promedio de las plantas de la parcela rociada con el producto es igual o no al de una parcela que no recibió el nuevo fertilizante. Se tomó una muestra de 20 brotes que tuvieron contacto con el fertilizante y 25 que no. Los datos obtenidos se muestran a continuación. Compruebe esa hipótesis con un nivel de significación de 5 por ciento:

	Muestra con fertilizante	Muestra sin fertilizante
Tamaño, n	$n_1 = 45$	$n_2 = 30$
Media	$\overline{X}_1 = 1.4$	$\overline{X}_2 = 0.75$
Desviación estándar	$S_1 = 0.3$	$S_2 = 0.5$

8. En un laboratorio de pruebas de calidad se desea examinar si la vida útil promedio (en días) de 2 marcas de focos de 100 watts es igual o no, con un nivel de significación de 0.05. Para ello, se tomó una muestra de 30 focos de cada marca, los resultados se muestran a continuación:

	Muestra marca A	Muestra marca B
Tamaño, n	$n_1 = 30$	$n_2 = 30$
Media	$\overline{X}_1 = 2017$	$\overline{X}_2 = 1950$
Desviación estándar	$S_1 = 173$	$S_2 = 121$

9. Una dependencia gubernamental encargada de supervisar cuestiones de salud pública desea probar si el contenido de nicotina de cigarros de una marca nacional y una extranjera son iguales o no, con un nivel de significación de 0.01. Se tomó una muestra de 50 cigarros de la marca nacional y 30 de la extranjera. Los datos obtenidos se muestran a continuación.

	Muestra marca extranjera	Muestra marca nacional
Tamaño, n	$n_1 = 30$	$n_2 = 50$
Media	$\overline{X}_1 = 15.8$	$\overline{X}_1 = 14.2$
Desviación estándar	$S_1 = 2.4$	$S_1 = 1.8$

10. Un consorcio industrial desea probar si el promedio de artículos defectuosos por cada lote de 1 000 difiere en 2 de sus plantas. En la planta A se tomó una muestra de 35 lotes y en la planta B de 32 lotes. Compruebe si existe diferencia con un nivel de significación de 2.5%. Los datos obtenidos se muestran a continuación.

	Muestra planta A	Muestra planta B
Tamaño, n	$n_1 = 35$	$n_2 = 32$
Media	$\overline{X}_1 = 13$	$\overline{X}_2 = 11$
Desviación estándar	$S_1 = 2.5$	$S_2 = 3$

10.2.2 Pruebas con muestras pequeñas e independientes, variables distribuidas normalmente

10.2.2.1 Cuando no se conocen las varianzas pero se asume que son iguales

11. En una construcción se emplean 2 tipos de bloque de hormigón, y se desea saber si la resistencia del bloque de hormigón tipo B es menor que la del A. Para investigarlo, del bloque A se tomó una muestra de 18 bloques y se encontró que tienen una resistencia media de 39.7 kg con una desviación estándar de 0.7 kg. Del bloque B se tomó una muestra de 14 y se encontró que tiene una resistencia media de 39.4 kg con una desviación estándar de 0.4 kg. Si la resistencia de los bloques se distribuye de forma normal y se sabe que sus varianzas son iguales, compruebe la hipótesis con un nivel de significación de 0.05.

12. Una muestra de 12 aspiradoras marca Bilmex demostró que en su operación gastan en promedio 44.84 kilowatt-hora con una desviación estándar de 8.7 kilowatt-hora. Mientras que una muestra de 17 aspiradoras marca Prix demostró que gastan en promedio 46.75 kilowatt-hora con una desviación estándar de 10.4 kilowatt-hora. Si el gasto de energía sigue una distribución normal y estudios anteriores demostraron que la varianza de las 2 marcas es igual compruebe la hipótesis de que el gasto de energía de las aspiradoras Bilmex es menor que el de las aspiradoras Prix con un nivel de significación de 0.025.

13. El gerente de personal de un restaurante desea evaluar si el promedio del monto de las propinas que reciben los meseros del turno vespertino es menor al del matutino, para lo cual tomó una muestra de 5 meseros de cada turno y registró cuánto recibieron de propina diariamente durante cierto número de días. Se encontró que los meseros del turno matutino recibieron en promedio \$632 con una desviación estándar de \$41.2; los meseros del turno vespertino recibieron en promedio \$685 con una desviación estándar de \$27.3. Si las propinas siguen una distribución normal y se sabe que las varianzas son iguales compruebe la hipótesis con un nivel de significación de 0.1.

14. Un gerente de compras está considerando 2 tipos de lámparas para el alumbrado de la planta ensambladora. Tomó una muestra de 8 lámparas de marca A y encontró que, en promedio, su tiempo de duración es de 1 455 horas con una desviación estándar de 87; de la lámpara de marca B tomó una muestra de 11 y encontró que en promedio duran 1 484 horas con una desviación estándar de 23. Si la vida útil de las lámparas sigue una distribución normal y se sabe que las varianzas son iguales compruebe la hipótesis de que las lámparas B duran menos que las A, con un nivel de significación de 0.005.

15. Se desea saber si el ingreso mensual, en miles, de los miembros de la asociación de contadores de la ciudad B es menor que el de la ciudad A. Para ello, de la ciudad A se tomó una muestra de 23 miembros y se encontró que en promedio ganan $32.52 mensuales con una desviación estándar de $5.48; mientras que en una muestra de 19 miembros de la ciudad B se encontró que en promedio ganan $24.67 con una desviación estándar de $4.38. Si el ingreso mensual sigue una distribución normal y se sabe que las varianzas son iguales compruebe la hipótesis con un nivel de significación de 0.05.

10.2.2.2 Cuando no se conocen las varianzas y no se asume que sean iguales

16. Del ejercicio 11, que trata sobre el uso de 2 tipos de bloque de hormigón de los que se desea saber si la resistencia de B es menor que la del A, se obtuvieron los siguientes datos:

	Bloque A	Bloque B
n	18	14
$\bar{X}$	39.7	39.4
s^2	0.007	0.004

Si no es posible asumir que las varianzas son iguales, compruebe la hipótesis con un nivel de significación del 0.05.

17. Se planea utilizar 2 tipos de baterías para un nuevo modelo de teléfono celular y se desea evaluar si la duración de una carga de la batería A es mayor que la de una de B, por lo que se tomaron muestras con los resultados siguientes:

	A	B
n	11	13
$\bar{X}$	38	42
s^2	4.5	5.1

Si no es posible asumir que las varianzas son iguales, demuestre la hipótesis con un nivel de significación de 0.025.

18. En el ejercicio 12, se deseaba probar la hipótesis de que el gasto de energía en la operación de las aspiradoras Bilmex es menor que el gasto de las aspiradoras Prix, y se obtuvieron los siguientes resultados de las muestras:

	Bilmex	Prix
n	12	17
$\bar{X}$	44.84	44.75
s^2	8.7	10.4

Si no es posible asumir que las varianzas son iguales, demuestre la hipótesis de que el gasto de energía de las aspiradoras Bilmex es menor que el de las aspiradoras Prix con un nivel de significación de 0.025.

19. Se quiere saber si el tiempo promedio que las niñas emplean en ver televisión diariamente es mayor al de los niños. Se tomó una muestra de 10 niños y 10 niñas, los resultados se muestran a continuación:

	Niños	Niñas
n	10	10
$\bar{X}$	5.7	4.8
s^2	2.2	1.9

Si no se puede suponer que las varianzas son iguales, compruebe la hipótesis con un nivel de significación de 0.1.

20. En el ejercicio 13, un gerente de personal desea demostrar si el promedio del monto de las propinas que reciben los meseros del turno vespertino es menor que el de los meseros del turno matutino, y se obtuvieron los siguientes datos:

	Matutino	Vespertino
n	5	5
$\bar{X}$	632	685
s^2	41.2	27.3

Si no se puede suponer que las varianzas son iguales, demuestre la hipótesis con un nivel de significación de 0.1.

10.2.3 Pruebas para muestras pareadas cuando no se conocen las varianzas pero no se necesita asumir que sean iguales

21. En una encuesta realizada a estudiantes de posgrado, una pregunta pedía asentar qué promedio general de calificaciones tenían en sus estudios, y para evaluar la exactitud de estas respuestas se decidió tomar una muestra de 12 de esos estudiantes para comparar sus

respuestas contra los datos que estaban anotados en los registros escolares. En la tabla siguiente se muestran los resultados que se obtuvieron:

Estudiante	Calificación según la encuesta	Calificación según los registros escolares
1	8.5	8.2
2	9.0	9.1
3	7.3	7.0
4	9.4	9.0
5	6.0	6.0
6	8.7	8.1
7	9.1	8.7
8	8.8	8.9
9	9.2	8.9
10	7.9	7.5
11	8.0	7.8
12	8.4	8.1

Compruebe si existe diferencia entre el promedio de calificaciones que los estudiantes respondieron en la encuesta y las que se tienen registradas en los archivos escolares, con un nivel de significación de 1 por ciento.

22. En una clínica de reducción de peso se afirma que su programa permite reducir en promedio más de 6 kg. En la tabla siguiente se muestra el resultado que obtuvieron 10 personas. Compruebe si la afirmación de la clínica es correcta, con un nivel de significación de 5 por ciento.

Cliente	Peso antes	Peso después
1	85.9	77.1
2	91.7	86.5
3	100.2	96.7
4	94.1	87.4
5	88.2	81.7
6	80.3	73.3
7	87.7	79.1
8	91.9	85.1
9	94.6	84.6
10	105.9	92.6

23. Para determinar la temperatura de la Tierra, se desea comparar las mediciones obtenidas a partir de termómetros en tierra contra las que arrojan termómetros aéreos, ya que ambos tipos trabajan en condiciones diferentes y cada uno tienen ventajas y desventajas propias. Se recogieron mediciones en 10 lugares diferentes con los 2 tipos de termómetros y se obtuvieron los resultados siguientes:

Lugar	Termómetro en tierra	Termómetro aéreo
1	46.8	47.2
2	45.5	48.2
3	36.2	37.8
4	31.1	32.8
5	24.6	26.1
6	22.4	23.4
7	49.7	50.1
8	40.6	42.7
9	37.6	39.3
10	35.6	38.0

Compruebe si existen diferencias entre los promedios de esas mediciones con un nivel de significación de 0.05 por ciento.

10.3 Pruebas de hipótesis sobre la diferencia entre 2 proporciones

En esta sección se revisan los procedimientos que deben utilizarse para realizar pruebas sobre la diferencia entre 2 proporciones cuando se tienen muestras independientes y tamaños de muestras grandes.

Cuando se desea probar una hipótesis sobre la diferencia entre 2 proporciones, puede emplearse la distribución normal si se tienen tamaños de muestra lo suficientemente grandes. El estadístico de prueba es:

$$z \cong \frac{p_1 - p_2}{p_c\left(1 - p_c\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \qquad \textbf{(10.13)}$$

En donde:

p_1 es la proporción de éxitos en la muestra 1.
p_2 es la proporción de éxitos en la muestra 2.
n_1 es el número de elementos de la muestra 1.

n_2 es el número de elementos de la muestra 2.
p_c es la combinación de las 2 proporciones, dado que la hipótesis nula supone que son iguales:

$$p_c = \frac{X_1 + X_2}{n_1 + n_1} \tag{10.14}$$

Aquí:

X_1 es el número de casos de éxito en la muestra 1.
X_2 es el número de casos de éxito en la muestra 2.

Repasando con esta última simbología, las anteriores p_1 y p_2 se calculan como:

$$p_1 = \frac{X_1}{n_1} \quad \text{y} \quad p_2 = \frac{X_2}{n_2}$$

De la expresión de z presentada antes puede verse que el error estándar de la diferencia entre 2 proporciones es, precisamente:

$$s_{p_1 - p_2} = p_c(1 - p_c)\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \tag{10.15}$$

EJEMPLO 10.6

Se desea probar si existe diferencia entre las proporciones de aficionados al futbol que tienen estudios universitarios y los que no. De una muestra aleatoria de 100 hombres con estudios universitarios se encuentra que 56 de ellos son aficionados a ese deporte. Otra muestra de 150 varones que no tienen esos estudios refleja que 90 manifiestan ser aficionados. Compruebe, con un nivel de significación de 0.05, si existe diferencia entre las proporciones de aficionados en ambas poblaciones.

Solución: Como se trata simplemente de probar si existe diferencia o no entre las 2 proporciones, se aplica una prueba de 2 extremos donde las hipótesis son:

$$H_0 : \pi_1 = \pi_2$$
$$H_1 : \pi_1 \neq \pi_2$$

Al ser $\alpha = 0.05$, y dado que se utilizará la distribución normal con una prueba de 2 extremos, se tiene que el valor crítico del estadístico de prueba es:

$$P(-1.96 \leq z \leq 1.96) = 0.05$$

Así, se rechazará H_0 si el valor de z calculado con los valores muestrales es menor que −1.96 o mayor que 1.96. El valor de la proporción combinada es:

$$p_c = \frac{X_1 + X_2}{n_1 + n_1} = \frac{56 + 90}{100 + 150} = \frac{146}{250} = 0.584$$

El valor calculado del estadístico de prueba es:

$$z \cong \frac{p_1 - p_2}{p_c(1 - p_c)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \frac{\frac{56}{100} - \frac{90}{150}}{0.584(0.416)\left(\frac{1}{100} + \frac{1}{150}\right)}$$

$$= \frac{0.56 - 0.60}{0.243(0.01 + 0.0067)} = \frac{-0.04}{0.0041} = -9.75$$

Como este valor calculado está muy por debajo de −1.96 resulta evidente que la diferencia entre los hombres sin estudios universitarios y los que sí los tienen es muy significativa, por lo que se rechaza la hipótesis nula.

EJEMPLO 10.7

En un proceso de producción se encontraron 35 artículos defectuosos dentro de una muestra aleatoria de 500, y se identificaron 20 defectuosos en otra muestra de 400 artículos provenientes de un proceso similar que se lleva a cabo en otra fábrica. Compruebe la hipótesis que afirma que los dos procesos producen la misma proporción de artículos defectuosos, con un nivel de significación de 1 por ciento.

Solución: Las hipótesis:

$$H_0 : \pi_1 = \pi_2$$
$$H_1 : \pi_1 \neq \pi_2$$

Como se trata de muestras grandes se puede utilizar la z, y con una prueba de 2 extremos y un nivel de significación de 0.01 se obtiene que el valor crítico de z es 2.575, ya que $P(-2.575 \leq z \leq 2.575) = 0.01$. La p combinada es:

$$p = \frac{x_1 + x_2}{n_1 + n_2} = \frac{35 + 20}{500 + 400} = 0.061$$

La z calculada con los datos muestrales es:

$$z = \frac{p_1 - p_2}{s_{(p_1 - p_2)}} = \frac{\frac{35}{500} - \frac{20}{400}}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.07 - 0.05}{\sqrt{0.061(0.939)\left(\frac{1}{500} + \frac{1}{400}\right)}} = \frac{0.02}{0.016} = 1.25$$

Debido a que la z calculada es mayor que -2.575 y menor que 2.575 no se puede rechazar la hipótesis nula, y se concluye que los 2 procesos producen la misma proporción de artículos defectuosos.

ejercicios 10.3 Pruebas de hipótesis sobre la diferencia entre 2 proporciones

Muestras independientes y tamaños de muestra grandes

1. Una agencia de colocación desea analizar las actitudes de los estudiantes de 2 universidades sobre la importancia del sueldo en el empleo, para lo cual realiza una encuesta a 300 estudiantes de una universidad particular y a 400 de una pública. En la privada 280 contestan que lo más importante en el empleo es el sueldo, en tanto que, en la pública 240 responden que el sueldo es lo más importante. Compruebe la hipótesis de que las proporciones de estudiantes de las 2 universidades que consideran el sueldo como lo más importante son iguales, con un nivel de significación de 1 por ciento.
2. Conforme se acerca el día de las elecciones en una importante ciudad, una empresa encuestadora revisa los resultados de 2 encuestas realizadas en fechas diferentes, y encuentra que en la primera encuesta en la que se entrevistó a 500 ciudadanos, 51% de ellos manifestaron estar a favor del candidato principal. En la encuesta posterior, realizada entre 800 electores, 47.5% manifestó estar a favor de ese candidato principal. Con un nivel de significación de 3%, ¿se puede afirmar que cambiaron las preferencias del electorado?
3. En un sondeo de opinión realizado por una estación de radio, 60 de 200 hombres dijeron que les gustaba determinado programa, y 75 de 300 mujeres manifestaron la misma opinión. Con un nivel de significación de 1%, ¿existe diferencia entre las proporciones de hombres y mujeres a los que les gusta ese programa de radio?
4. Un gerente de finanzas está analizando el comportamiento de sus cuentas por pagar, ha obtenido muestras de cuentas del mes de mayo de 2 años consecutivos. En mayo del año 1, con una muestra de 1 300 cuentas por pagar, descubrió 50 que no habían sido liquidadas en el plazo convenido, mientras que en una muestra de 1 000 cuentas de mayo del año 2, había 38 que no se pagaron a tiempo. Con un nivel de significación de 5%, ¿se puede afirmar que ha habido un aumento en la proporción de cuentas por pagar que caen en la morosidad?
5. Se aplicaron cambios en una línea de producción a fin de reducir el porcentaje de artículos defectuosos. Para evaluar si las modificaciones resultaron efectivas, se tomaron muestras de 300 artículos antes y después de realizar los cambios, y se encontró que antes el porcentaje de defectuosos era de 3% y después resultó de 4%. Compruebe si las modificaciones redujeron la proporción de artículos defectuosos, con un nivel de significación de 2 por ciento.

10.4 Prueba para la diferencia entre 2 varianzas

Para probar si existe o no diferencia entre las varianzas de 2 poblaciones puede utilizarse como estadístico de prueba la F de la distribución *F de Fisher*, llamada así en honor del destacado estadístico Ronald Aylmer Fisher, que se calcula como el cociente de las varianzas de las 2 poblaciones:

$$F = \frac{\sigma_1^2}{\sigma_2^2} \tag{10.16}$$

Que sería la expresión teórica de F. Y el valor calculado de F a partir de las varianzas muestrales:

$$F_{cal} = \frac{s_1^2}{s_2^2} \tag{10.17}$$

La prueba se lleva a cabo sobre la diferencia hipotética entre 2 varianzas poblacionales: $H_0 : \sigma_1^2 - \sigma_2^2 = 0$; para realizarla se obtienen las varianzas de 2 muestras tomadas de 2 poblaciones diferentes. En otras palabras, esta prueba se realiza para poblaciones independientes, a las que suele identificarse como 1 y 2. Las 2 varianzas muestrales son las que se utilizan como base para hacer inferencias sobre sus correspondientes parámetros.

Si puede asumirse que las 2 varianzas poblacionales son iguales, $\sigma_1^2 = \sigma_2^2$, entonces se utiliza, como estadístico de prueba, la distribución F con $n_1 - 1$ grados de libertad para el numerador y $n_1 - 1$ grados de libertad para el denominador; ya que el estadístico de prueba se calcula con los datos muestrales se construye con un cociente, como se verá.

La distribución F no es una distribución simétrica; está sesgada a la derecha y su forma específica depende de los grados de libertad tanto del numerador como del denominador. En la tabla 4 del apéndice de tablas se muestran algunos valores de probabilidad seleccionados para esta distribución. Nótese que la tabla está dividida en 3 secciones: para 0.05, 0.025 y 0.01, áreas de probabilidad en el extremo derecho de la distribución; es decir, en realidad son 3 tablas, una para cada una de las áreas mencionadas. A su vez, cada tabla tiene como encabezados de las columnas los grados de libertad del numerador y en los renglones, los grados de libertad del denominador. Así, para un área de 0.05 en el extremo derecho de esta distribución, con 10 grados de libertad en el numerador y 20 en el denominador, el valor de F es igual a 2.35. Este valor quiere decir que, dados esos grados de libertad, la probabilidad de que la F sea igual o mayor a 2.35 es de 0.05, o de 5%. Esto mismo en símbolos:

$$P(F \geq 2.35 \mid gl_1 = 10, gl_2 = 20) = 0.05$$

Tal como puede apreciarse, al tratarse de una distribución asimétrica, la tabla de la distribución F no muestra valores de probabilidad para el lado izquierdo y éstos se requieren cuando la prueba que se realiza es de 2 extremos ($\neq$) o cuando es de un extremo y la región de rechazo está en el lado izquierdo. En estos casos, para determinar los valores no mostrados, lo que se hace es utilizar el inverso del valor correspondiente de las tablas, invirtiendo el orden de los grados de libertad. En símbolos:

$$F_{1-\alpha, gl_2, gl_1} = \frac{1}{F_{\alpha, gl_2, gl_1}} \tag{10.18}$$

Se ilustran los conceptos anteriores con algunos ejemplos; en la sección 10.4.1 sobre la distribución y Excel se explicará la forma en la que este paquete de Microsoft permite construir la tabla 4 del apéndice y, en general, obtener cualquier valor del estadístico de prueba F de Fisher.

EJEMPLO 10.8

Se desea comparar el grado de aprendizaje en matemáticas en 2 escuelas del mismo nivel que utilizan métodos de enseñanza diferentes. Para aplicar la prueba t para la diferencia entre 2 medias, debe ser posible suponer que ambas poblaciones tienen la misma varianza. Por ello, antes de realizar la prueba sobre las medias, es conveniente realizar una prueba sobre la igualdad de varianzas de las 2 poblaciones. Al hacer esta prueba, se toma una muestra aleatoria de 21 estudiantes en cada una de las 2 escuelas y se obtienen los siguientes resultados:

Escuela 1	Escuela 2
$n_1 = 21$	$n_2 = 21$
$\bar{x} = 7.9$	$\bar{x} = 8.3$
$s_1 = 1.1$	$s_1 = 1.21$

Se desea realizar la prueba de hipótesis con $\alpha = 0.05$. Las hipótesis son:

$$H_0 : \sigma_1^2 - \sigma_2^2 = 0$$
$$H_1 : \sigma_1^2 - \sigma_2^2 = 0$$

Como se trata de una prueba de 2 extremos, se divide el nivel de significación en 2, por lo que los valores críticos del estadístico de prueba son:

Para el extremo derecho:

$$P(F \geq 2.46 \mid gl_1 = 20, gl_2 = 20) = 0.025$$

En símbolos:

$$F_{0.025, 20, 20} = \frac{1}{2.46} = 0.41$$

Para el extremo izquierdo:

$$P(F < 0.41 \mid gl_1 = 20, gl_2 = 20) = 0.025$$

Nótese que en este caso no fue necesario intercambiar los valores de los grados de libertad, ya que son los mismos en ambas muestras. Ahora se calcula el estadístico F a partir de los valores muestrales, la F empírica:

$$F_{emp} = \frac{s_1^2}{s_2^2} = \frac{1.1^2}{1.21^2} = \frac{1.21}{1.4641} = 0.83$$

Entonces, como este valor empírico de F es mayor que 0.41 y menor que 2.46, no se rechaza la hipótesis nula y se concluye que ambas poblaciones de estudiantes tienen la misma varianza.

A su vez, esta conclusión permite decidir que sí es adecuada la prueba t para probar la hipótesis de la igualdad de las medias de las 2 poblaciones, como sigue:

$$H_0 : \overline{X}_1 - \overline{X}_2 = 0$$
$$H_1 : \overline{X}_1 - \overline{X}_2 \neq 0$$

Si se asume que las varianzas de las 2 poblaciones son iguales, se combinan las 2 varianzas muestrales:

$$s_c^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{20(1.1^2) + 20(1.21^2)}{21 + 21 - 2} =$$

$$\frac{24.2 + 29.282}{40} = 1.4641$$

Y el error estándar de la diferencia entre 2 medias:

$$s_{\overline{X}_1 - \overline{X}_2} = \sqrt{\frac{s_c^2}{n_1} + \frac{s_c^2}{n_2}} = \sqrt{\frac{1.4641}{21} + \frac{1.4641}{21}}$$

$$= \sqrt{0.06972 + 0.06972} = 0.1394$$

El valor del estadístico de prueba t:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_{\overline{X}_1 - \overline{X}_2}} = \frac{7.9 - 8.3}{0.1394} = \frac{-0.4}{0.1394} = -2.87$$

Como el valor crítico de t es 2.021 dado que:

$$P(-2.021 \leq t \leq 2.021) \mid gl = 40) = 2.021$$

Entonces, se rechaza H_0 y se concluye que la diferencia entre las 2 medias muestrales es diferente de cero.

Ejemplo 10.9

En las bolsas de valores resulta de interés comparar 2 varianzas como en el caso de los precios de los títulos que se negocian: acciones, bonos o certificados, entre otros. La variabilidad medida por la varianza (o la desviación estándar, que es la raíz cuadrada de la varianza) mide la dispersión de los datos alrededor de la media y suele utilizarse como medida del riesgo de los títulos; se asume que a mayor variabilidad, es decir a mayor varianza o desviación estándar, mayor es el riesgo.

Al comparar el riesgo de las acciones de AMXL (telefónica América Móvil) y las de GeoB (de la inmobiliaria Casas Geo) se supone, de acuerdo con las características de estas 2 empresas, que son más riesgosas las acciones de Geo que las de AMX por lo que se creería que la varianza de las acciones de AMX es menor que la de las acciones de Geo. Para evaluar esta situación, se tomaron muestras de 42 días de precios de las acciones de cada una de estas empresas que cotizan en la Bolsa Mexicana de Valores, y se obtuvieron varianzas de 2.79 y 4.11, en ese orden. Compruebe, con un nivel de significación de 1%, si existe una diferencia real entre estas 2 varianzas.

Solución: Las hipótesis son:

$$H_0 : \sigma_1^2 - \sigma_2^2 = 0$$
$$H_1 : \sigma_1^2 - \sigma_2^2 < 0$$

O alternativamente:

$$H_0 : \sigma_1^2 = \sigma_2^2$$
$$H_1 : \sigma_1^2 < \sigma_2^2$$

En donde el subíndice 1 identifica a AMX y el 2 a Geo.
Como se trata de una prueba del extremo izquierdo, se encuentra que para el extremo derecho:

$$P(F \geq 2.094 \mid gl_1 = 41, gl_2 = 41) = 0.01$$

Y para el extremo izquierdo, utilizando la fórmula (10.18):

$$F_{1-\alpha, gl_2, gl_1} = \frac{1}{F_{\alpha, gl_2, gl_1}} = \frac{1}{2.094} = 0.478$$

El estadístico F a partir de los valores muestrales, la F empírica:

$$F_{emp} = \frac{s_1^2}{s_2^2} = \frac{1.1^2}{1.21^2} = \frac{2.79}{4.11} = 0.679$$

Entonces, como este valor empírico de F es mayor que 0.478, no se rechaza la hipótesis nula y se concluye que ambas acciones tienen la misma varianza y, por lo tanto, el mismo riesgo.

10.4.1 Distribución *F* y Excel

Este paquete de Microsoft tiene 2 funciones que permiten calcular valores de la distribución F de Fisher:

DISTR.F(x, grados_de_libertad_1,grados de libertad2)
DISTR.F.INV(probabilidad,grados_de_libertad_1,grados de libertad2)

La primera produce la probabilidad, es decir el área, que el valor de F señalado por x aísla en el extremo derecho de la distribución para los valores anotados de los grados de libertad 1 (del numerador) y los grados de libertad 2 (del denominador) y para un nivel de significación de 10%. No es de mucha utilidad para los propósitos de este libro, ya que está restringido a ese nivel de significación.

Por su parte, la función Distr.F.Inv arroja el valor de F que divide la probabilidad registrada en la parte derecha de la distribución para los grados de libertad anotados, que son, precisamente, los valores de F obtenidos en la tabla 4 del apéndice.

Para ilustrar este mecanismo, en el ejemplo 10.8 se buscaba el valor de F que aislara el 0.025 del área en la parte derecha de la distribución, con 20 grados de libertad tanto en el numerador como en el denominador. Si se anota en alguna celda de Excel la función =DISTR.F.INV(0.025,20,20) se obtiene el valor 2.464484299, el mismo que resultó en ese ejemplo a partir de las tablas sólo que con mayor precisión.

En el ejemplo 10.9 se buscaba el valor de F que aislara el 0.01 del área en la parte derecha de la distribución, con 40 grados de libertad tanto en el numerador como en el denominador. Si se anota en alguna celda de Excel la función =DISTR.F.INV(0.01,40,40) se logra el valor 2.114232454, igual al que se obtuvo en ese ejemplo a partir de las tablas, pero más preciso.

Con esta función de Excel se elaboró la tabla 4 del apéndice de tablas, la tabla de los valores de F para diferentes áreas en el extremo derecho de la distribución y para diferentes combinaciones de grados de libertad para el numerador y el denominador.

En vista de lo anterior, puede concluirse que en toda ocasión posible es más preciso utilizar esta función de Excel para obtener con exactitud los valores de F que se requieran; sin embargo, cuando esto no sea posible o necesario, pueden emplearse los valores de la tabla 4 a pesar de que estén limitados por cuestiones de espacio. Utilizando la función de Excel pueden determinarse valores de F para cualquier valor de probabilidad y para cualquier combinación de grados de libertad.

EJERCICIOS 10.4 Prueba para la diferencia entre 2 varianzas

1. Un fabricante de luminarias sospecha que una de sus líneas de producción está fabricando productos con demasiada variabilidad en su vida útil. Para evaluar la situación, decide comparar la varianza de una muestra de 25 de esas luminarias contra la varianza de otra muestra de 30 procedentes de una línea de producción que no está bajo sospecha. Se determinan las correspondientes varianzas y se encuentra que la de la línea de producción sospechosa es de 900, en tanto que la otra es de 350. ¿Puede concluirse, con un nivel de significación de 1%, que la variabilidad de la vida útil de las luminarias fabricadas en la línea de producción bajo sospecha es mayor que la de aquellas que provienen de la línea de producción que trabaja adecuadamente?
2. Veintiún personas de entre 20 y 35 años de edad que participaron en un curso de capacitación basado en un texto programado, promediaron 58.22 min para terminar el curso, con una varianza de 8.5 min. En tanto, en un curso simultáneo de personas mayores de 35 años, el promedio fue de 63.24 min con una varianza de 9.2 min. Compruebe la hipótesis de que la varianza de los 2 cursos es la misma con un nivel de significación de 5 por ciento.
3. Supervisores experimentados hacen la varianza de las mediciones del grosor de la cubierta de 41 protectores de plástico registrando que es de 1.39 cm. Si un supervisor recién contratado mide una muestra de 25 cubiertas y encuentra que tienen una varianza de 1.47 cm compruebe la hipótesis de que el nuevo supervisor está haciendo mediciones correctas, con un nivel de significación de 1 por ciento.
4. Una muestra aleatoria de 31 cajas de mangos de la carga de un proveedor arrojó una varianza de 2.5 kg para su peso. Una muestra de 27 cajas de mangos de otro proveedor reflejó una varianza de 4.2 kg. ¿Puede aceptarse la hipótesis de que la varianza de los pesos de las 2 cargas es la misma? Realice la prueba con un nivel de significación de 1 por ciento.
5. Un comentarista deportivo desea saber si la varianza del tiempo de recorrido del maratón de la Ciudad de México ha cambiado desde las últimas mediciones que indican, según una muestra aleatoria de 61 corredores, una varianza de una hora. Tome una muestra aleatoria de 30 corredores y observe una varianza de 0.87 de hora. Realice la prueba con un nivel de significación de 0.05.

10.5 Excel y pruebas de hipótesis para 2 muestras

Este paquete de Microsoft tiene 2 tipos de herramientas que pueden utilizarse en el tema de pruebas de hipótesis para 2 poblaciones. Por un lado, tiene la función de la Prueba.F, y por el otro tiene 5 herramientas integradas en el complemento de "Análisis de datos":

- Prueba F para varianzas de 2 muestras.
- Prueba t para medias de 2 muestras emparejadas (pareadas).
- Prueba t para medias de 2 muestras suponiendo varianzas iguales.
- Prueba t para medias de 2 muestras suponiendo varianzas desiguales.
- Prueba z para medias de 2 muestras.

La función Prueba.F puede emplearse para probar la igualdad de las varianzas de 2 poblaciones a partir de los datos de 2 muestras. Su sintaxis es:

PRUEBA.F(matriz1,matriz2)

Se obtiene como resultado la probabilidad de una sola cola de que las varianzas de las 2 muestras de datos no sean significativamente diferentes.

Tanto la función Prueba.F —que sirve para probar la igualdad de varianzas de 2 poblaciones— como las 5 pruebas —incluidas en la opción "Análisis de datos"— se aplican a partir de los dos conjuntos de datos muestrales. Todos los ejemplos desarrollados aquí fueron resueltos partiendo de que ya se conocían las medidas muestrales (medias y/o desviaciones estándar), así que no es posible ilustrar estas herramientas de Excel con ellos, puesto que no se dispone de los valores muestrales a partir de los cuales se calcularon esas medidas muestrales.

Entonces, no se abundará más sobre estas herramientas, sin embargo se sugiere tenerlas presentes para cuando sea necesario resolver pruebas de hipótesis y se cuente con los datos de las muestras.

10.6 Resumen

En este capítulo se estudió la metodología básica necesaria al realizar pruebas de hipótesis para las medidas correspondientes a 2 poblaciones, y se revisaron las pruebas para la diferencia entre 2 medias en diversas circunstancias:

1. Con muestras grandes e independientes, cuando se conocen y cuando no se conocen las varianzas correspondientes a las 2 poblaciones. Además se explicaron 2 casos para esta última circunstancia, cuando no se conocen las varianzas:
   a) puede asumirse que son iguales,
   b) no puede asumirse que lo sean.
2. Pruebas para 2 poblaciones con muestras pequeñas e independientes, variables distribuidas normalmente, cuando no se conocen las varianzas de las correspondientes poblaciones:
   a) puede asumirse que son iguales,
   b) no puede asegurarse que lo sean.
3. Pruebas para muestras pareadas cuando no se conocen las varianzas pero no se necesita asumir que sean iguales.
4. Prueba para la diferencia entre 2 proporciones.
5. Prueba para la diferencia entre 2 varianzas.

Asimismo, en la sección 10.1 se presentó un panorama general de las pruebas de hipótesis, en donde se incluyen los temas presentados en este capítulo. También se adelantó una revisión a los temas de pruebas de hipótesis que se abordarán en capítulos posteriores, una visión de conjunto de este importante tema de estadístico.

Del mismo modo se ha explicado la manera de utilizar las funciones de Excel en este tema de pruebas para 2 poblaciones, así como el procedimiento para construir una tabla de valores de la distribución *F* de Fisher, que se emplea tanto aquí como en temas posteriores.

10.7 Fórmulas del capítulo

En seguida se desarrollan las fórmulas elaboradas en cada sección de este capítulo.

10.2.1 Para la diferencia entre 2 medias con muestras grandes e independientes

El estadístico de prueba (la *z* estandarizada o valor calculado de *z*) cuando se conocen las varianzas de las 2 poblaciones:

$$z = \frac{\bar{X}_1 - \bar{X}_2}{\sigma_{X_1 - X_2}} \qquad (10.1)$$

El error estándar de la diferencia entre 2 medias es:

$$\sigma_{X_1 - X_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \qquad (10.2)$$

El estadístico de prueba (la *z* estandarizada o valor calculado de *z*) cuando no se conocen las varianzas de las 2 poblaciones:

$$z = \frac{\bar{X}_1 - \bar{X}_2}{s_{X_1 - X_2}} \qquad (10.3)$$

El error estándar de la diferencia entre 2 medias cuando no se conocen las varianzas de las poblaciones:

$$s_{X_1 - X_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \qquad (10.4)$$

10.2.1.2 Cuando no se conocen las varianzas pero se asume que son iguales

Cálculo de la varianza combinada a partir de las varianzas muestrales para el cálculo del error estándar:

$$s_c^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \qquad (10.5)$$

10.2.2 Pruebas con muestras pequeñas e independientes, variables distribuidas normalmente

Error estándar de la diferencia de medias cuando no se conocen las varianzas poblacionales pero se supone que son iguales:

$$s_{X_1 - X_2} = \sqrt{\frac{s_c^2}{n_1} + \frac{s_c^2}{n_2}} \quad (10.6)$$

10.2.2 Estadístico *t* para muestras pequeñas e independientes, no se conocen las varianzas poblacionales y se asume que la variable se distribuye de forma normal en las 2 poblaciones:

$$t = \frac{\left(\bar{X}_1 - \bar{X}_2\right)}{s_{x_1} - x_2} \quad (10.7)$$

Con $n_1 + n_2 - 2$ grados de libertad.

Estadístico *t* modificado para cuando se tienen muestras pequeñas e independientes, se puede asumir que la variable se distribuye de forma normal en las 2 poblaciones, no se conocen las varianzas poblacionales pero no se puede asumir que sean iguales:

$$t'_{cr} = \frac{\frac{s_1^2}{n_1} t_1 + \frac{s_2^2}{n_2} t_2}{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \quad (10.8)$$

10.2.3 Pruebas para muestras pareadas cuando no se conocen las varianzas pero no se necesita asumir que sean iguales

Promedio de las diferencias entre observaciones pareadas, con muestras pequeñas; se desconocen las varianzas pero no se asume que sean iguales y la variable se distribuye de forma normal en la población:

$$\bar{D} = \frac{\sum D_i}{n} \quad (10.9)$$

La varianza de las diferencias pareadas:

$$S = \sqrt{\frac{\sum_{i=1}^{n}\left(D_i - \bar{D}\right)^2}{n - 1}} \quad (10.10)$$

El error estándar de las diferencias pareadas:

$$s_D = \frac{s}{\sqrt{n}} \quad (10.11)$$

El estadístico de prueba:

$$t_{n-1} = \frac{\bar{D}}{s_D} \quad (10.12)$$

Con $n - 1$ grados de libertad.

10.3 Pruebas de hipótesis sobre la diferencia entre 2 proporciones

El estadístico de prueba para hipótesis sobre la diferencia entre 2 proporciones con muestras independientes y tamaños de muestra grandes:

$$Z \cong \frac{p_1 - p_2}{p_c\left(1 - p_c\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \quad (10.13)$$

La combinación de las 2 proporciones, dado que la hipótesis nula supone que son iguales:

$$p_c = \frac{x_1 + x_2}{n_1 + n_1} \quad (10.14)$$

El error estándar de la diferencia entre 2 proporciones:

$$s_{p_1 - p_2} = p_c\left(1 - p_c\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \quad (10.15)$$

10.4 Prueba para la diferencia entre 2 variables

La distribución de probabilidad *F* de Fisher, teórica:

$$F = \frac{\sigma_1^2}{\sigma_2^2} \quad (10.16)$$

La distribución de probabilidad *F* de Fisher, empírica:

$$F_{emp} - \frac{s_1^2}{s_2^2} \quad (10.17)$$

$$F_{1-\alpha, gl_2, gl_1} = \frac{1}{F_{\alpha, gl_2, gl_1}} \quad (10.18)$$

10.8 Ejercicios adicionales

10.2.1 Pruebas con muestras grandes e independientes

10.2.1.2 Cuando no se conocen las varianzas y no se asume que sean iguales

1. En un estudio se desea conocer si existe diferencia entre las horas promedio de sueño de los estudiantes de licenciatura de 2 universidades. En la universidad A se tomó una muestra de 43 estudiantes y se encontró que en promedio duermen 5.7 horas diarias con una desviación estándar de 1.4; mientras que en la universidad B, en una muestra de 40 alumnos se encontró que en promedio duermen 4.8 horas con una desviación estándar de 1.04. Compruebe la hipótesis con un nivel de significación de 5 por ciento.
2. Se tomó una muestra de 35 alumnos de 6o. de primaria de 2 escuelas diferentes. A cada alumno se le dio un texto de 3 párrafos para leerlo en voz alta mientras se le tomaba el tiempo. Para la escuela A el tiempo promedio de lectura fue de 4.7 minutos con una desviación estándar de 0.94; mientras que en la escuela B el tiempo promedio fue de 4.9 minutos con una desviación estándar de 1.35. Compruebe la hipótesis de que no existe diferencia entre el tiempo promedio de lectura de los alumnos de las 2 escuelas, con un nivel de significación de 0.01.
3. Se tomó una muestra de 37 reproductores de sonido marca Sintac, y otra de 35 marca Tonx, se midió el nivel de decibeles que producen al subir el volumen al máximo. En promedio, los reproductores Sintac registraron 157 decibeles con una desviación estándar de 8.3; y los reproductores Tonx 148 decibeles con una desviación estándar de 6.5. ¿Puede asegurarse que no existe diferencia entre los promedios de decibeles de ambas marcas? Demuestre la hipótesis con un nivel de significación de 4 por ciento.

4. Se desea conocer si existe diferencia entre el número de pacientes que se reciben en los servicios de emergencias de 2 hospitales de la misma zona. Se tomó una muestra de 31 días en cada hospital, en el A en promedio se recibieron 42 pacientes con una desviación estándar de 8, y en el hospital B se recibieron en promedio 53 pacientes con una desviación estándar de 10. Compruebe la hipótesis con un nivel de significación de 2.5 por ciento.
5. Se realizó una prueba que midió la velocidad, para descargar archivos de poco peso, en 2 diferentes servidores de internet. En una prueba de 40 descargas para el servidor A y 48 descargas para el servidor B se encontró que, en promedio, el servidor A descarga un archivo en 3.1 minutos con una desviación estándar de 0.14, mientras que para el servidor B el promedio fue de 2.8 minutos con una desviación estándar de 0.2. Demuestre la hipótesis de que no existe diferencia entre la velocidad de descarga de los 2 servidores, con un nivel de significación de 5 por ciento.

10.2.1.3 Cuando no se conocen las varianzas pero se asume que son iguales

6. Se desea probar, con un nivel de significación de 5%, si existe diferencia entre los tiempos de absorción de 2 jarabes para la tos. Para ello, se tomó una muestra de 55 personas que tomaron el jarabe A y 50 que tomaron el B y se obtuvieron los datos que se muestran a continuación:

	Muestra jarabe A	Muestra jarabe B
Tamaño, n	$n_1 = 55$	$n_2 = 50$
Media	$\overline{X}_1 = 67$	$\overline{X}_2 = 53$
Desviación estándar	$S_1 = 3.8$	$S_2 = 2.7$

7. Se está evaluando el voltaje (máximo) de operación de los circuitos manufacturados por 2 fabricantes. Para probar si existe diferencia en los voltajes de operación en los circuitos de los 2 fabricantes, se tomó una muestra de 45 del circuito A y 30 del B; en un laboratorio de calidad los circuitos fueron sometidos a pruebas de voltaje, y se obtuvieron los siguientes resultados acerca de su operación:

	Muestra circuito A	Muestra circuito B
Tamaño, n	$n_1 = 45$	$n_2 = 30$
Media	$\overline{X}_1 = 67$	$\overline{X}_2 = 53$
Desviación estándar	$S_1 = 3.8$	$S_2 = 2.7$

Determine si la afirmación es correcta con un nivel de significación de 4.5 por ciento.

8. Un fabricante de impresoras de inyección de tinta maneja 2 modelos: deluxe y simple. Se desea evaluar si existe diferencia en el rendimiento promedio de los cartuchos de impresión y se tomó una muestra de 35 cartuchos del modelo simple y 33 del de lujo. Los datos obtenidos se muestran a continuación:

	Muestra de lujo	Muestra simple
Tamaño, n	$n_1 - 33$	$n_2 - 35$
Media de hojas impresas	$\overline{X}_1 = 564$	$\overline{X}_1 = 537$
Desviación estándar	$S_1 = 18$	$S_2 = 27$

Determine si la afirmación es correcta con un nivel de significación de 4 por ciento.

9. En una fábrica de productos de cartón que cuenta con 2 secciones de armado, el gerente de personal desea probar, con un nivel de significación de 3%, si existe diferencia entre la productividad de los trabajadores en términos de las cajas que arman en cada periodo de 2 horas. Se tomó una muestra de 38 periodos de cada sección y se obtuvieron los siguientes datos:

	Muestra sección A	Muestra sección B
Tamaño, n	$n_1 = 38$	$n_1 = 38$
Promedio de cajas armadas por periodo	$\overline{X}_1 = 1\,524$	$\overline{X}_1 = 1\,496$
Desviación estándar	$S_1 = 8$	$S_2 = 6$

10.2.2 Pruebas con muestras pequeñas e independientes, variables distribuidas normalmente

10.2.2.1 Cuando no se conocen las varianzas pero se asume que son iguales

10. Una empresa del ramo alimenticio desea probar que en el centro B se atienden menos llamadas que en el A. Para ello se tomó una muestra de 10 días para cada centro y se encontró que, en promedio, en el centro A se atienden 219 llamadas diarias con una desviación estándar de 32; mientras que en el centro B se atienden 197 llamadas con una desviación estándar de 19. Si las llamadas recibidas siguen una distribución normal y las varianzas son iguales compruebe la hipótesis con un nivel de significación de 0.1.
11. En 2 centros educativos de nivel básico se ha implementado un nuevo método para la enseñanza de matemáticas. Se quiere demostrar que, luego de un examen diagnóstico, el promedio obtenido por los estudiantes de la escuela A es menor que el de los estudiantes de la B. Para ello se tomó una muestra de 15 estudiantes de cada escuela y se les aplicó el examen. Se encontró que en la escuela A el promedio fue de 7.8 con una desviación estándar de 0.4; en la escuela B el promedio fue de 8.2 con una desviación estándar de 0.6. Si las calificaciones siguen una distribución normal y se sabe que las varianzas son iguales compruebe la hipótesis con un nivel de significación de 0.025.
12. El gerente encargado de 2 plantas maquiladoras desea conocer si existe diferencia en el promedio de prendas que son devueltas por algún defecto en cada lote de 500 piezas, esto entre los pedidos surtidos en cada una de las maquilas. Para ello se tomó una muestra de 12 lotes de la planta norte y se encontró que en promedio fueron devueltas 17 piezas

con una desviación estándar de 5; mientras que en la planta centro se tomó una muestra de 15 lotes y se descubrió que en promedio se regresan 14 piezas con una desviación estándar de 7. Compruebe la hipótesis de que el número de piezas devueltas por defecto de los lotes surtidos por la planta norte es menor que las devueltas en los lotes de la planta centro, con un nivel de significación de 0.1.

13. Se desea conocer si la reacción de las mujeres a cierto estímulo es menor que la de los hombres. Para ello se tomó una muestra de 9 mujeres donde la reacción promedio en milisegundos era de 483 con una desviación estándar de 5.4; mientras que en una muestra de 7 hombres el promedio de reacción fue de 467 con una desviación estándar de 7.3. Compruebe la hipótesis con un nivel de significación de 0.05.

14. La Secretaría de Agricultura reparte 2 tipos de fertilizantes a pequeños productores; desea probar si el crecimiento de las cosechas donde se emplea el fertilizante B es menor que el de las cosechas en que se usa el tipo A. Así que se tomó una muestra de 7 campos de siembra del mismo tamaño para cada tipo de fertilizante. En los que se emplea el fertilizante A, el crecimiento promedio fue de 184 cm con una desviación estándar de 4.7, y para los que utilizan el fertilizante B el crecimiento promedio fue de 174 cm con una desviación estándar de 3.2. Compruebe la hipótesis con un nivel de significación de 0.01.

10.2.2.2 Cuando no se conocen las varianzas y no se asume que sean iguales

15. Para llevar a cabo cierto proceso pueden utilizarse 2 tipos de máquinas. El departamento de control ambiental quiere demostrar que las emisiones de gases en gramos de la máquina B es menor que la de A; se midió la cantidad de gases emitidos por cada máquina en 7 repeticiones del proceso. Los resultados se muestran a continuación:

	Máquina A	Máquina B
n	7	7
$\overline{X}$	47.8	52
s^2	3.6	5.1

Si no puede suponerse que las varianzas son iguales, demuestre la hipótesis con un nivel de significación de 0.005.

16. Un gerente de adquisiciones está considerando 2 tipos de lámparas para el alumbrado de la planta ensambladora. Para probar si existe diferencia en la duración promedio de dichos objetos, se tomó una muestra de 8 lámparas de tipo A y 11 de tipo B, y se obtuvieron los siguientes datos para el promedio de duración en horas:

	Lámpara A	Lámpara B
n	8	11
$\overline{X}$	1 455	1 484
s^2	87	23

Si no puede suponerse que las varianzas son iguales, demuestre la hipótesis de que las lámparas B duran más que las A, con un nivel de significación del 0.005.

17. En un laboratorio se está haciendo un estudio que compara la cantidad de agua, en ml, necesaria para que una semilla orgánica y una transgénica germinen. Se tomó una muestra de 15 semillas orgánicas y 17 transgénicas, los resultados son los siguientes:

	Orgánica	Transgénica
n	15	17
$\overline{X}$	577	458
s^2	48	32

Si no puede suponerse que las varianzas sean iguales, compruebe la hipótesis de que la cantidad de agua requerida para que una semilla transgénica germine es menor que la de una orgánica, con nivel de significación de 0.05.

18. Un gremio de contadores desea conocer si el ingreso mensual, en miles de pesos, de los miembros de la asociación de la ciudad B es mayor que el de los miembros de la A. Se tomaron muestras en ambas ciudades y se obtuvieron los siguientes datos:

	Ciudad A	Ciudad B
n	23	19
$\overline{X}$	32.52	24.67
s^2	5.48	4.38

Si no puede suponerse que las varianzas sean iguales, demuestre la hipótesis con un nivel de significación de 0.05.

19. El gerente de producción de una empresa desea demostrar que el promedio de defectos, en el armado de estuches para cosméticos, de la división norte es menor al de la división centro. Para ello, tomó una muestra de 12 lotes de cada división y encontró los resultados siguientes:

	Norte	Centro
n	12	12
$\overline{X}$	27	32
s^2	5	3

Si no es posible asumir que las varianzas son iguales, compruebe la hipótesis con un nivel de significación de 0.05.

10.2.3 Pruebas para muestras pareadas cuando no se conocen las varianzas pero no se necesita asumir que sean iguales

20. Una empresa de logística con problemas en la venta de servicios, debido al nivel de uso de tiempo extra utilizado en la terminación de los mismos, comienza un programa para mejorar la calidad. Se toma una muestra de 12 clientes que informan sobre las horas de tiempo extra empleadas

en la finalización de los servicios, antes y después del programa implementado. Los directivos quieren saber si hubo un cambio en el uso de tiempo extra en la conclusión de los servicios, teniendo en cuenta un nivel de significación de 0.05.

Cliente	Horas de tiempo extra después del programa	Horas de tiempo extra antes del programa
1	6.8	7.1
2	7.3	7.5
3	8.1	8.1
4	7.1	7.2
5	6.3	6.7
6	8.4	9.1
7	8.1	8.6
8	7.6	7.6
9	6.1	6.4
10	7.3	7.6
11	7.8	8.1
12	6.7	7.1

21. El área de proyectos de inversión de una empresa tiene los recursos excedentes en 8 fondos de inversión de los cuales desea obtener mejores rendimientos, así que implementa durante 10 meses 5 cursos de capacitación para los empleados que los administran, con lo cual obtiene los siguientes rendimientos:

Fondo	Rendimiento anterior %	Rendimiento posterior %
1	8.7	8.9
2	12.4	11.6
3	21.3	21.5
4	6.4	6.4
5	8.1	8
6	9.4	9.1
7	18.3	18.5
8	13.7	13.8

El gerente de finanzas desea saber si hubo un incremento en el rendimiento de los fondos utilizando un nivel de significación de 0.025.

22. Una maquiladora de playeras promocionales está teniendo muchas devoluciones por fallas de confección; a fin de disminuir las devoluciones, el nuevo jefe de producción establece ciertas medidas de control de calidad para ser ejecutadas por los trabajadores, además de invertir en el mantenimiento correctivo de las máquinas. Se hace una comparación, de los 18 meses anteriores con los 18 meses posteriores a los cambios, para ver si disminuyeron las devoluciones. Utilice un nivel de significación de 0.01.

Meses	Núm. de devoluciones antes de los cambios	Núm. de devoluciones después de los cambios
1	425	420
2	376	380
3	645	601
4	728	690
5	397	413
6	566	521
7	476	501
8	758	397
9	636	610
10	487	505
11	689	605
12	743	690
13	650	550
14	546	610
15	723	701
16	679	660
17	574	551
18	677	523

23. Una empresa que se dedica a la venta de productos por catálogo desea incrementar sus ventas, entonces ofrece a sus vendedores 1.5% más de comisión sobre las ventas que realicen. Una muestra aleatoria de 14 trabajadores señala la siguiente información:

Vendedor	Ventas sin comisión adicional ($)	Ventas con 1.5% adicional de comisión ($)
1	6 750	7 200
2	8 200	9 570
3	12 987	12 800
4	5 370	6 650
5	7 800	8 300
6	3 020	4 540
7	16 970	17 280
8	12 500	12 290
9	9 780	8 900
10	7 550	8 750
11	6 900	7 300
12	13 430	14 130
13	5 500	6 270
14	9 890	9 150

Determine, con un nivel de significación de 0.05, si hubo incremento en las ventas después de ofrecer mayor porcentaje de comisión a los vendedores.

24. Una tienda de renta de videojuegos compara el número de suscriptores mensuales durante 6 meses antes y 6 meses después de que la competencia llegara al D.F., y obtiene los siguientes resultados:

Mes	Antes de la competencia	Después de la competencia
1	73	65
2	104	92
3	98	88
4	123	110
5	145	121
6	163	127

Determine si el número de suscripciones mensuales se vio afectado por la llegada de la competencia al D.F., teniendo en cuenta un nivel de significación de 0.025.

10.3 Pruebas de hipótesis sobre la diferencia entre 2 proporciones

Muestras independientes y tamaños de muestra grandes

25. Conforme se acerca el día de las elecciones en una importante ciudad, una empresa encuestadora revisa los resultados de 2 encuestas realizadas en fechas diferentes, y encuentra que en la primera encuesta en la que se entrevistó a 500 ciudadanos, 50.4% de ellos manifestaron estar a favor del candidato principal. En la encuesta posterior, realizada entre 800 electores, el 47.5% manifestó estar a favor de ese candidato principal. Con un nivel de significación de 3%, ¿puede afirmarse que cambiaron las preferencias del electorado?

26. En un sondeo de opinión realizado por una estación de radio, 15 de 50 hombres dijeron que les gustaba determinado programa, mientras que 20 de 75 mujeres manifestaron la misma opinión. Con un nivel de significación de 1%, ¿existe diferencia entre las proporciones de hombres y mujeres a los que les gusta ese programa de radio?

27. Un gerente de finanzas está analizando el comportamiento de sus cuentas por pagar, ha obtenido muestras de cuentas del mes de mayo de 2 años consecutivos. En mayo del año 1, con una muestra de 1 300 cuentas por pagar, descubrió 50 que no habían sido liquidadas en el plazo convenido, mientras que en una muestra de 1 000 cuentas de mayo del año 2, había 50 que se pagaron a tiempo. Con un nivel de significación de 5%, ¿puede afirmarse que ha habido un aumento en la proporción de cuentas por pagar que caen en la morosidad?

28. Se aplicaron cambios en una línea de producción a fin de reducir el porcentaje de artículos defectuosos. Para evaluar si las modificaciones resultaron efectivas, se tomaron muestras de 100 artículos antes y después de realizar los cambios, y se encontró que antes el porcentaje de defectuosos era de 4% y después resultó de 3%. Compruebe si las modificaciones redujeron la proporción de artículos defectuosos, con un nivel de significación de 2 por ciento.

10.4 Prueba para la diferencia entre 2 varianzas

29. Un estudio de investigación obtiene la siguiente información: en una muestra aleatoria de 21 empleados del sector comercial se cumplen con 85 h, en promedio, de capacitación al año con una varianza de 4.5 h. Mientras que una muestra aleatoria de 30 empleados del sector educativo demuestra que cumplen con 135 h promedio de capacitación anualmente con una varianza de 5.6 h. Determine si la varianza de las horas promedio de capacitación al año que toman los 2 sectores es la misma, teniendo en cuenta un nivel de significación de 0.05.

30. Dos ensambladoras de automóviles toman una muestra aleatoria de 11 coches cada una, la primera empresa encuentra que en su muestra la vida útil de los autos es de 17 años en promedio con una varianza de 3.2 años. La segunda empresa encuentra que en su muestra el promedio es de 19 años con una varianza de 2.7 años. Determine, con un nivel de significación de 0.02, si las varianzas son iguales.

31. Las principales máquinas de una fábrica están automáticamente programadas para que después de cierta producción y desgaste se detengan el tiempo necesario para continuar produciendo. Una muestra aleatoria de 9 máquinas de la planta 1 demuestra que trabajan 63 h continuas con una varianza de 0.3 h, mientras que una muestra aleatoria de 7 máquinas de la planta 2 demuestra que trabajan en promedio 65 h con una varianza de 3.3 h. Determine, con un nivel de significación de 1%, si la varianza es la misma en las 2 plantas o si debe haber un cambio en la programación de éstas.

32. Un entrenador de natación desea saber si la varianza de la velocidad promedio de los nadadores de 2 equipos es la misma. Una muestra aleatoria de 10 nadadores del primer equipo demuestra que en promedio su velocidad es de 1.6 m/s, con una varianza de 0.02 m/s. La muestra aleatoria del mismo número de nadadores del segundo equipo refleja que su velocidad promedio es de 1.56 m/s con una varianza de 0.025 m/s. Determine si hay diferencia entre las varianzas considerando un nivel de significación de 5 por ciento.

33. Una agencia de viajes quiere saber si los guías turísticos nativos tienen la misma varianza en el tiempo de recorrido, de cierto lugar, que los guías foráneos. Una muestra aleatoria de 61 recorridos de los guías nativos afirma que en promedio les toma 5.5 h llevarlo a cabo con una varianza de 1.7 h. La muestra aleatoria de 81 recorridos de los guías foráneos demuestra que éstos tardan en promedio 7 h en hacer el mismo recorrido con una varianza de 3.2 h. Determine, con un nivel de significación de 1%, si la varianza es la misma.

capítulo 11

Pruebas de hipótesis con la distribución ji cuadrada

Sumario

11.1 Introducción
11.2 Distribución ji cuadrada χ^2
11.3 Tablas de áreas bajo la curva de la distribución ji cuadrada
 11.3.1 Excel y la tabla de áreas para χ^2
11.4 Pruebas de hipótesis para la varianza de una población
11.5 Distribución ji cuadrada a partir de frecuencias observadas y frecuencias esperadas
11.6 Pruebas para una proporción con z y con χ^2
 11.6.1 Prueba de una proporción con z
 11.6.2 Prueba de una proporción con χ^2
11.7 Prueba para la diferencia entre 2 proporciones con z y con χ^2
 11.7.1 Prueba para la diferencia entre 2 proporciones con z
 11.7.2 Prueba para la diferencia entre 2 proporciones con χ^2
11.8 Relación entre las pruebas de hipótesis para proporciones con z y con χ^2
11.9 Prueba para la diferencia entre n proporciones
11.10 Pruebas de bondad de ajuste a distribuciones teóricas
 11.10.1 Pruebas de bondad de ajuste a una distribución normal
 11.10.2 Pruebas de bondad de ajuste a una distribución Poisson
 11.10.3 Pruebas de bondad de ajuste a una distribución binomial
11.11 Pruebas de bondad de ajuste entre distribuciones empíricas
11.12 Pruebas sobre la independencia entre 2 variables
11.13 Pruebas paramétricas y pruebas no paramétricas
11.14 Excel y la distribución ji cuadrada
 11.14.1 Función Distr. Chi
 11.14.2 La función Prueb. Chi
11.15 Resumen
11.16 Fórmulas del capítulo
11.17 Ejercicios adicionales

11.1 Introducción

χ^2 (ji cuadrada)[1] se utiliza para realizar diversas pruebas de hipótesis. Las que se revisan aquí son:

- Prueba para una varianza.
- Prueba para una proporción.
- Prueba para la diferencia entre 2 proporciones.
- Prueba para la diferencia entre n proporciones.
- Prueba de bondad de ajuste.
- Prueba de independencia.
- Prueba de homogeneidad.

En las secciones siguientes se revisarán las particularidades de cada una de estas pruebas de hipótesis, ahora sólo se desea resaltar que la primera prueba se lleva a cabo utilizando la varianza de la muestra, en tanto que las 6 restantes se realizan comparando frecuencias observadas y frecuencias esperadas. A su vez, en estas pruebas con frecuencias observadas y esperadas (en las que se refieren a una y a 2 proporciones) se muestra cómo las pruebas con χ^2 son equivalentes a las pruebas con el estadístico normal z, las cuales se estudiaron en capítulos anteriores.

[1] Es frecuente ver en libros de texto que a esta distribución se le llama *chi cuadrada* lo cual es un error debido a una mala traducción al inglés del nombre de esa letra griega, ya que ése es su nombre textual en ese idioma. El nombre correcto en español es *ji cuadrada*.

En la sección siguiente se revisarán las principales características de la distribución χ^2, incluyendo su función de densidad. En el punto 11.3 se describirá el uso de la tabla de áreas bajo la curva de la distribución ji cuadrada. En el apartado 11.4 se ilustrará el procedimiento para realizar pruebas de hipótesis sobre una varianza poblacional, mientras que en el 11.5 se demostrará que puede plantearse la distribución ji cuadrada como el cuadrado del estadístico z de la distribución normal estándar, y la misma distribución χ^2 obtenida a partir de frecuencias observadas y esperadas, que resulta ser la base para las secciones 11.6 a 11.12. Estos últimos subcapítulos explicarán las pruebas de hipótesis para la diferencia entre una y 2 proporciones, sobre la diferencia entre n proporciones y pruebas de bondad de ajuste, tanto de independencia como de homogeneidad.

11.2 Distribución ji cuadrada (χ^2)

Se recordará, del capítulo 6, las distribuciones de probabilidad continuas, en donde se explicó que si X es una variable aleatoria que se distribuye de manera normal con media μ y desviación estándar σ, entonces la variable, que se muestra a continuación, se distribuye también de forma normal con media 0 y desviación estándar 1.

$$z = \frac{X - \mu}{\sigma}$$

Esta distribución se conoce como *distribución normal estándar*, y que se tabula en las tablas de áreas bajo la curva normal, se ha utilizado con frecuencia en capítulos anteriores. Se tiene que el cuadrado de esta variable z^2 es:

$$z^2 = \frac{(X - \mu)^2}{\sigma^2}$$

A partir de z^2 puede obtenerse la variable u para una muestra de tamaño n, mediante la sumatoria de z^2 para todos los elementos de la muestra:

$$u = \sum_{i=1}^{n} \left[\frac{(X_i - \mu)^2}{\sigma^2} \right]$$

Si ahora se obtiene la distribución muestral de u, es decir, el conjunto de las u de todas las muestras posibles que pueden extraerse de tamaño n, provenientes de la correspondiente población de tamaño N, la distribución que se obtiene es lo que se conoce como *distribución χ^2 para n grados de libertad*. Se ilustra ahora su función de densidad de probabilidad:

$$f(u) = \frac{1}{\left(\frac{n}{2} - 1\right)! 2^{\frac{n}{2}}} u^{\frac{n}{2} - 1} e^{-\frac{u}{2}}, \qquad u > 0 \tag{11.1}$$

En donde:

$$u = \sum_{i=1}^{n} \left[\frac{(X_i - \mu_i)^2}{\sigma_i^2} \right] \tag{11.2}$$

$e = 2.71828$.
n = número de observaciones.

Dado que existe una distribución χ^2 para cada número posible de grados de libertad, entonces existe una cantidad ilimitada de distribuciones χ^2, ya que los grados de libertad pueden tomar valores desde 1 hasta infinito.

En las pruebas de hipótesis, el número de grados de libertad está dado por $n - k - 1$ en donde k es el número de parámetros que se estiman (en los ejemplos siguientes se ilustra este concepto con mayor detenimiento). Si se utiliza gl para representar los grados de libertad, lo anterior se expresa como:

$$gl = n - k - 1$$

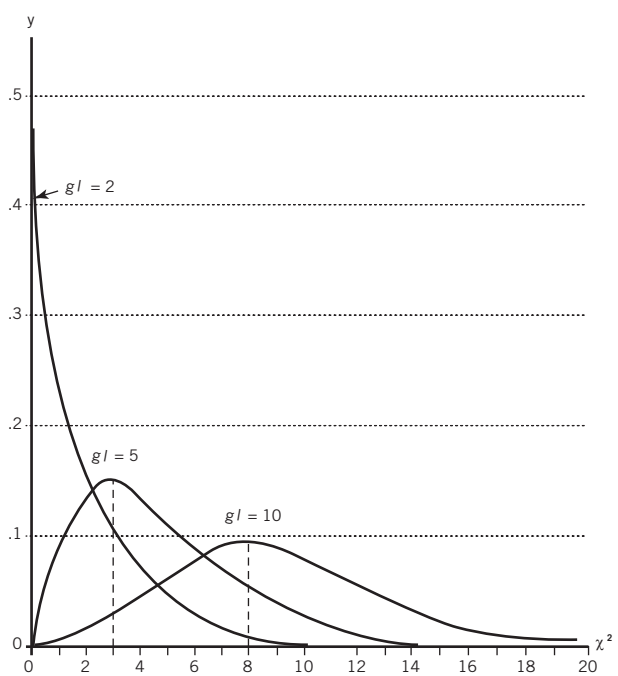


Figura 11.1 Distribución χ^2 para diversos grados de libertad.

En la figura 11.1 se muestran las gráficas de esta distribución para diversos valores de los grados de libertad. En esta figura pueden distinguirse algunas de las principales características de la distribución ji cuadrada:

- Tiende a la simetría conforme aumentan los grados de libertad.
- No puede tener valores por debajo de cero (debido a que se obtiene de números elevados al cuadrado).
- La moda de la distribución es igual al número de grados de libertad menos 2, $gl - 2$. Por ejemplo, la moda para esta distribución cuando $gl = 10$, es igual a 8, tal como puede apreciarse en la gráfica.
- El área bajo la curva es igual a 100 por ciento.
- El eje y muestra las probabilidades.

11.3 Tablas de áreas bajo la curva de la distribución ji cuadrada

No es sencillo determinar las áreas bajo la curva tanto de la distribución normal como de otras distribuciones, entonces suelen utilizarse tablas para facilitar su determinación.

En la tabla 11.1 se presentan áreas bajo la curva de la distribución χ^2 para los valores de probabilidad que se utilizan con mayor frecuencia. Asimismo, se incluyen grados de libertad de 1 a 40, uno por uno, y de 40 a 100 de 10 en 10. Esta tabla corresponde a la número 5 del "Apéndice de tablas" incorporado al final del texto.

Tabla 11.1 Áreas bajo la curva de la distribución χ^2

	Probabilidad (porción de área en el extremo derecho)					
gl/prob.	0.500	0.100	0.050	0.025	0.010	0.005
1	0.455	2.706	3.841	5.024	6.635	7.879
2	1.386	4.605	5.991	7.378	9.210	10.597
3	2.366	6.251	7.815	9.348	11.345	12.838
4	3.357	7.779	9.488	11.143	13.277	14.860
5	4.351	9.236	11.070	12.833	15.086	16.750
6	5.348	10.645	12.592	14.449	16.812	18.548
7	6.346	12.017	14.067	16.013	18.475	20.278
8	7.344	13.362	15.507	17.535	20.090	21.955
9	8.343	14.684	16.919	19.023	21.666	23.589
10	9.342	15.987	18.307	20.483	23.209	25.188
11	10.341	17.275	19.675	21.920	24.725	26.757
12	11.340	18.549	21.026	23.337	26.217	28.300
13	12.340	19.812	22.362	24.736	27.688	29.819
14	13.339	21.064	23.685	26.119	29.141	31.319
15	14.339	22.307	24.996	27.488	30.578	32.801
16	15.338	23.542	26.296	28.845	32.000	34.267
17	16.338	24.769	27.587	30.191	33.409	35.718
18	17.338	25.989	28.869	31.526	34.805	37.156
19	18.338	27.204	30.144	32.852	36.191	38.582
20	19.337	28.412	31.410	34.170	37.566	39.997
21	20.337	29.615	32.671	35.479	38.932	41.401
22	21.337	30.813	33.924	36.781	40.289	42.796
23	22.337	32.007	35.172	38.076	41.638	44.181

	Probabilidad (porción de área en el extremo derecho)					
gl/prob.	0.500	0.100	0.050	0.025	0.010	0.005
24	23.337	33.196	36.415	39.364	42.980	45.559
25	24.337	34.382	37.652	40.646	44.314	46.928
26	25.336	35.563	38.885	41.923	45.642	48.290
27	26.336	36.741	40.113	43.195	46.963	49.645
28	27.336	37.916	41.337	44.461	48.278	50.993
29	28.336	39.087	42.557	45.722	49.588	52.336
30	29.336	40.256	43.773	46.979	50.892	53.672
31	30.336	41.422	44.985	48.232	52.191	55.003
32	31.336	42.585	46.194	49.480	53.486	56.328
33	32.336	43.745	47.400	50.725	54.776	57.648
34	33.336	44.903	48.602	51.966	56.061	58.964
35	34.336	46.059	49.802	53.203	57.342	60.275
36	35.336	47.212	50.998	54.437	58.619	61.581
37	36.336	48.363	52.192	55.668	59.893	62.883
38	37.335	49.513	53.384	56.896	61.162	64.181
39	38.335	50.660	54.572	58.120	62.428	65.476
40	39.335	51.805	55.758	59.342	63.691	66.766
50	49.335	63.167	67.505	71.420	76.154	79.490
60	59.335	74.397	79.082	83.298	88.379	91.952
70	69.334	85.527	90.531	95.023	100.425	104.215
80	79.334	96.578	101.879	106.629	112.329	116.321
90	89.334	107.565	113.145	118.136	124.116	128.299
100	99.334	118.498	124.342	129.561	135.807	140.169

Por ejemplo, el número 39.997 que está en el extremo derecho del renglón 20, y que corresponde a una probabilidad de 0.005, señala que la probabilidad de que χ^2 sea mayor o igual que 39.997, con 20 grados de libertad, es de 0.005 o 0.5%. Lo mismo en simbología de probabilidad:

$$P(\chi^2 \geq 39.997 \mid gl = 20) = 0.005$$

11.3.1 Excel y la tabla de áreas para χ^2

La tabla 11.1 se construyó utilizando la función de Excel PRUEBA.CHI.INV (probabilidad, grados de libertad) la cual proporciona el valor de ji cuadrada para la probabilidad y los grados de libertad especificados.

Siguiendo con la misma ilustración del párrafo anterior, al introducir en Excel la función = PRUEBA.CHI.INV(0.005, 20) se obtiene el mismo valor de 39.997.

Resulta evidente que con esta función de Excel puede determinarse el valor crítico de χ^2 para cualquier par de circunstancias (probabilidad y grados de libertad) que no estén contemplados en la tabla 11.1.

11.4 Pruebas de hipótesis para la varianza de una población

Al recordar que la varianza y su raíz cuadrada, la desviación estándar, miden dispersión con respecto a la media, logra apreciarse por qué se pueden utilizar pruebas sobre estas medidas cuando se requiere probar la uniformidad o variabilidad de algún proceso o producto. Por ejemplo, si se desea verificar si cierto tipo de cristal es lo suficientemente homogéneo como para utilizarlo en la fabricación de equipo óptico de precisión; al desear evaluar si el diámetro de cierta tubería está dentro de límites especificados.

Las pruebas sobre una varianza poblacional son aplicables solamente cuando la variable se distribuye de manera normal en la población y la hipótesis nula es de la siguiente forma:

$$H_0: \sigma^2 = \sigma_0^2$$

En donde σ_0^2 es el valor que se supone tiene la verdadera varianza poblacional. En estas pruebas se utiliza como estadístico de prueba la ji cuadrada, χ^2, que se calcula como:

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} \tag{11.3}$$

Tiene $n - 1$ grados de libertad. Aquí, se reducen en 1 los grados de libertad porque no se estima ningún parámetro. Se revisa en seguida un ejemplo.

EJEMPLO 11.1

En la fabricación de cierto tipo de tubo de acero se requiere que la varianza del peso no exceda a 5.3 g^2. Si una muestra aleatoria de 30 tubos tiene una varianza de 8.25 g al cuadrado, con un nivel de significación de 1%, ¿puede concluirse a partir de estos datos que la norma se cumple?

Solución: Las hipótesis

$$H_0: \sigma^2 \leq 5.3$$
$$H_1: \sigma^2 > 5.3$$

El nivel de significación: $\alpha = 0.01$, de donde:

$$P(\chi^2 \geq 49.588 \mid gl = 29) = 0.01$$

El valor calculado del estadístico de prueba:

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{8.25(30-1)}{5.3} = 45.14$$

Dado que el valor calculado del estadístico de prueba es menor que el del crítico, no se rechaza la hipótesis nula y se concluye que el proceso cumple con la norma de la variabilidad del peso de los tubos.

EJEMPLO 11.2

Cuando el proceso está bajo control, el diámetro del cuerpo de un tornillo tiene una varianza de 0.000127 cm^2. Para evaluar si el proceso está en esas condiciones, se toma una muestra aleatoria de 25 de esos tornillos y se obtiene una varianza de 0.000155 cm^2.

a) Con un nivel de significación del 0.05 ¿puede concluirse que el proceso está bajo control?

b) ¿En qué suposición se basa la respuesta al inciso *a*)?

Solución:

a) Las hipótesis

$$H_0: \sigma^2 = 0.000127$$
$$H_1: \sigma^2 \neq 0.000127$$

El nivel de significación: $\alpha = 0.05$ de donde:

$$P(\chi^2 \geq 36.415 \mid gl = 24) = 0.05$$

El valor calculado del estadístico de prueba:

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{0.000155(25-1)}{0.000127} = 29.29$$

Como el valor calculado del estadístico de prueba es menor que el crítico, no se rechaza la hipótesis nula y se concluye que el proceso está bajo control.

b) La respuesta se basa en la suposición de que el diámetro del cuerpo de los tornillos se distribuye de manera normal.

ejercicios 11.4 Pruebas de hipótesis para la varianza de una población

1. Próximamente se evaluará la efectividad en el proceso de embotellamiento de una bebida con presentación de 1 L, el procedimiento tiene una varianza de 0.000037 mL^2. Como medida preventiva se quiere revisar que este indicador no haya cambiado, por lo que se toma una muestra aleatoria de 81 botellas del producto y se obtiene que la varianza es de 0.000095 con un nivel de significación de 0.025, ¿puede asegurarse que la varianza es la misma?
2. En una distribuidora se tenía establecido que cada trabajador debía empacar 85 cajas en 90 minutos con una varianza de 6; al inicio de este periodo se impartieron 2 cursos de capacitación para reducir los costos del tiempo de empaque, por lo cual se espera que cada trabajador empaque la misma cantidad de cajas en menos tiempo, y en consecuencia que la varianza también disminuya. Una muestra aleatoria de 65 trabajadores señala que en promedio tardan 78 minutos con una varianza de 4 en empacar las 85 cajas. Determine con un nivel de significación de 0.1 si debe aceptarse la hipótesis de que la varianza en los tiempos se redujo.
3. Una agencia de viajes tiene un precio estándar para ir a varios destinos que estén ubicados a menos de 1 000 km de la capital, y que la varianza de esos viajes sea inferior a 64 km^2. Uno de los nuevos vendedores asegura que la agencia tiene mal categorizados los destinos, por lo que se toma una muestra aleatoria de 35 viajes que arroja una varianza de 65 km^2. Con un nivel de significación de 0.1, ¿debe aceptarse la hipótesis del vendedor, es decir, reconocer que el cálculo está mal hecho?
4. Una empresa que se dedica a la reparación de aparatos eléctricos registró hace 6 meses que, en promedio, gasta \$120 en cada trabajo con una desviación estándar de \$5. El presupuesto del próximo año lo realizará un despacho independiente y necesita verificar los datos, así que toma una muestra aleatoria de 40 trabajos cuya desviación estándar es de \$4.90. Con un nivel de significación de 0.01, ¿puede aceptarse que la varianza del costo de la reparación siga siendo la misma?
5. Una comercializadora de agua embotellada tiene establecido que la varianza del contenido de sodio en su presentación de 1.5 L debe ser inferior a 0.0020 mg^2. Una muestra aleatoria de 20 botellas de esta presentación refleja que la varianza es de 0.0025. Determine con un nivel de significación de 0.005 si realmente se está cumpliendo con lo establecido.

11.5 Distribución ji cuadrada a partir de frecuencias observadas y frecuencias esperadas

En esta sección se muestra cómo puede convertirse la expresión anterior de la distribución ji cuadrada, desprendida de:

$$u = \sum_{i=1}^{n} \left[\frac{(X_i - \mu_i)^2}{\sigma_i^2} \right]$$

Para transformarse en la expresión que se usa en las pruebas de hipótesis, a revisar en las secciones siguientes, que es:

$$\chi^2 = \sum \left[\frac{(O_i - E_i)^2}{E_i} \right]$$

En donde:

O_i son las frecuencias observadas.
E_i son las frecuencias esperadas.

Si se supone un experimento binomial y se sustituyen los subíndices i por e para *éxito* y f para *fracaso*, se tiene:

$$\sum \left[\frac{(O_i - E_i)^2}{E_i} \right] = \frac{(O_e - E_e)^2}{E_e} + \frac{(O_f - E_f)^2}{E_f}$$

Ahora, como E_e es la frecuencia esperada de éxitos y se trata de una variable binomial, de acuerdo con lo revisado en el capítulo 5 para esta distribución, su media es $\mu = np$, por lo que puede escribirse la expresión anterior como:

$$\frac{(O_e - E_e)^2}{E_e} + \frac{(O_f - E_f)^2}{E_f} = \frac{(O_e - np)^2}{np} + \frac{(O_f - E_f)^2}{E_f}$$

Dado que las frecuencias observadas de fracaso son el número de casos n, menos las frecuencias observadas de éxito, se traduce en símbolos:

$$O_f = n - O_e$$

Esta última expresión:

$$\frac{(O_e - np)^2}{np} + \frac{(O_f - E_f)^2}{E_f}$$

Puede escribirse como:

$$\frac{(O_e - np)^2}{np} + \frac{\left[(n - O_e) - E_f\right]^2}{E_f}$$

Ya que las frecuencias esperadas de fracaso también son el número de casos n, menos las frecuencias esperadas de éxito, se puede escribir:

$$E_f = n - E_e = n - np = n(1 - p)$$

Por lo que la expresión anterior quedaría representada como:

$$\frac{(O_e - np)^2}{np} + \frac{\left[(n - O_e) - n(1 - p)\right]^2}{n(1 - p)}$$

Dado que:

$$[(n - O_e) - n\,(1 - p)]^2 = (n - O_e - n - np)^2 = (-O_e + np)^2 = (O_e - np)^2$$

La expresión anterior se reduce a:

$$\frac{(O_e - np)^2}{np} + \frac{(O_e - np)^2}{n(1 - p)}$$

La cual, a su vez, sumando y simplificando:

$$\frac{(O_e - np)^2}{np} + \frac{(O_e - np)^2}{n(1 - p)} = \frac{(1 - p)(O_e - np)^2 + p(O_e - np)^2}{np(1 - p)}$$

$$= \frac{(O_e - np)^2\left[(1 - p) + p\right]}{np(1 - p)} = \frac{(O_e - np)^2}{np(1 - p)}$$

Esta última expresión también puede plantearse como:

$$\frac{(O_e - np)^2}{np(1 - p)} = \frac{(O_e - E_e)^2}{npq}$$

Tal como se explicó en el capítulo 5, npq es la varianza de la distribución binomial, O_e es el número de éxitos observados y E_e es la media, por lo que:

$$\frac{(O_e - E_e)^2}{npq} = \frac{(X - \mu)^2}{\sigma^2}$$

La misma expresión de donde se partió.

Así, se obtiene que, para muestras grandes la expresión es:

$$\chi^2 = \sum \left[\frac{(O_i - E_i)^2}{E_i} \right] \tag{11.4}$$

Se distribuye aproximadamente como ji cuadrada, χ^2, con grados de libertad $gl = n - k - 1$ que varían según el número de observaciones, n, y según la cantidad de parámetros que se estimen, k. En los ejemplos que se proporcionarán en las secciones siguientes se muestra cómo determinar estos grados de libertad de acuerdo con diferentes circunstancias.

11.6 Pruebas para una proporción con z y con χ^2

Aquí se explica cómo puede realizarse una prueba de hipótesis sobre la proporción de una población mediante 2 métodos equivalentes que, por supuesto, conducen a la misma conclusión. En primer lugar, se reproduce un ejemplo revisado en el capítulo 9 que trata sobre pruebas de hipótesis para una población; en segundo, se resuelve el mismo ejemplo pero ahora utilizando esta distribución χ^2. Finalmente se resalta que se llega a la misma conclusión y se analiza la relación entre ambos métodos.

11.6.1 Prueba de una proporción con z

Se reproduce como ejemplo 11.3 el ejemplo 9.12 que ilustra una prueba de hipótesis para la proporción de una población.

Ejemplo 11.3

El coordinador de la bolsa de trabajo de una universidad pública afirma que al menos 30% de los alumnos que terminan sus estudios obtiene empleo antes de 3 meses. Para probar esta afirmación, se toma una muestra de 50 estudiantes de dicha institución y se encuentra que sólo 10 obtuvieron empleo durante los primeros 3 meses luego de haber terminado sus estudios. ¿Puede rechazarse la afirmación de ese coordinador, con un nivel de significación de uno por ciento?

Solución: Como se trata de una muestra grande puede utilizarse z como estadístico de prueba.

Las hipótesis:

$$H_0: \pi = 0.30$$
$$H_1: \pi \neq 0.30$$

Dado el planteamiento que conduce a estas hipótesis se sabe que se trata de una prueba de 2 extremos y como $\alpha = 0.01$ se divide esta probabilidad entre los 2 extremos de la curva normal; se obtiene que los valores de z que dividen las regiones de aceptación y de rechazo son: $-2.575 \leq z \leq 2.575$. En la figura 11.2 se ilustra esta información.

El error estándar de la proporción:

$$\sigma_p = \sqrt{\frac{\pi Q}{n}} = \sqrt{\frac{0.3(0.7)}{50}} = \sqrt{0.0042} = 0.065$$

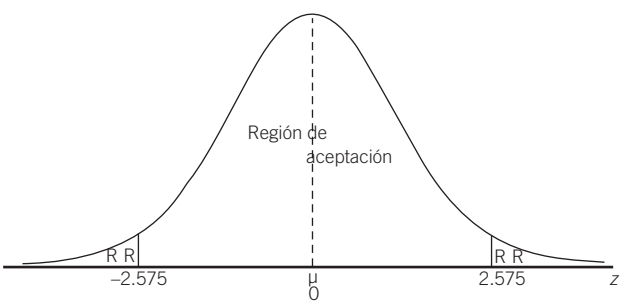


Figura 11.2 Condiciones del ejemplo 11.3.

La proporción de la muestra es $p = 11/50 = 0.20$, por lo que:

$$z = \frac{p - \pi}{\sigma_p} = \frac{p - \pi}{\sqrt{\frac{\pi Q}{n}}} = \frac{0.20 - 0.30}{0.065} = \frac{0.10}{0.065} = 1.54$$

Debido a que este valor del estadístico de prueba calculado con los datos muestrales es mayor que $z = -2.575$ y menor que $z = 2.575$, no es posible rechazar la hipótesis nula; se concluye que la proporción de estudiantes que terminan sus estudios y que consiguen un empleo antes de 3 meses sigue siendo de 30 por ciento.

11.6.2 Prueba de una proporción con χ^2

Se resuelve el ejemplo anterior pero utilizando la ji cuadrada como estadístico de prueba.

EJEMPLO 11.4

El coordinador de la bolsa de trabajo de una universidad pública afirma que al menos 30% de los alumnos que terminan sus estudios obtiene empleo antes de 3 meses. Para probar esta afirmación, se toma una muestra de 50 estudiantes de dicha institución y se encuentra que sólo 10 obtuvieron empleo durante los primeros 3 meses luego de haber terminado sus estudios. ¿Puede rechazarse la afirmación de ese coordinador, con un nivel de significación de uno por ciento?

Solución: Las hipótesis siguen siendo las mismas:

$$H_0: \pi = 0.30$$
$$H_1: \pi \neq 0.30$$

	Alumnos graduados		
	Con empleo	Sin empleo	Total
f_o	10	40	50
f_e	15	35	50

En este caso los grados de libertad son: $gl = k - m - 1 = 2 - 0 - 1 = 1$. La tabla con las frecuencias de los que tienen empleo (p) y los que no lo tienen (q, o sea $1 - p$) se trata de una tabla de 2×2 (con empleo, sin empleo y f_o, f_e) por lo que también puede aplicarse la regla para determinar los grados de libertad. Una tabla como ésta, conocida como *tablas de contingencias*, indica que los grados de libertad se calculan como: $gl = (c - 1)(r - 1)$. En donde c es el número de columnas y r es el número de renglones. Por ello $gl = (2 - 1)(2 - 1) = 1$, al igual que cuando se aplica el otro criterio.

Entonces el valor crítico del estadístico de prueba es:

$$P(\chi^2 \geq 7.87 \mid gl = 1) = 0.005$$

El valor calculado de χ^2 es:

$$\chi^2 = \frac{(10 - 15)^2}{15} + \frac{(40 - 35)^2}{35} = 1.67 + 0.714 = 2.38$$

Ya que el valor calculado de ji cuadrada es menor que el valor crítico, no se rechaza la hipótesis nula y se concluye que la afirmación del coordinador es correcta: el porcentaje de estudiantes que terminan sus estudios y que tienen trabajo antes de 3 meses es de 30%. Por supuesto, esta conclusión es la misma a la que se llegó mediante la prueba con la z de la distribución normal como estadístico de prueba.

EJERCICIOS 11.6 Pruebas para una proporción con z y χ^2

1. Una página de internet, reconocida mundialmente, patrocina una campaña que afirma que al menos 50% de los oficinistas navegan por internet. Para probar esta afirmación se toma una muestra de 120 trabajadores y se encuentra que 50 de ellos usan este servicio. ¿Puede rechazarse la afirmación de la campaña con un nivel de significación de uno por ciento?
2. Una empresa dedicada a fabricar televisores afirma que 40% de sus consumidores prefieren las televisiones de 32 pulgadas. Para comprobar esta afirmación, se toma una muestra de 80 clientes y se determina que 50 de ellos prefieren las televisiones de 32 pulgadas. ¿Puede rechazarse la afirmación de la empresa fabricadora de televisores, con un nivel de significación de 0.8 por ciento?
3. Una empresa cigarrera afirma que como mínimo 60% de la población adulta ha consumido su producto. Para comprobar esta afirmación, se toma una muestra de 120 personas y 80 de ellos afirman haber consumido esta marca. ¿Puede rechazarse la afirmación de la empresa, con un nivel de significación de 2 por ciento?
4. Un grupo de empresas de telefonía móvil afirma que 40% de sus clientes prefieren los teléfonos de prepago que los de contrato. Para probar esto se toma una muestra de 200 clientes y se encuentra que 110 prefiere el servicio de prepago. ¿Puede rechazarse la afirmación del grupo de empresas, con un nivel de significación de uno por ciento?
5. Un banco nacional afirma que no más de 80% de los trabajadores del país cuenta con una tarjeta, ya sea de crédito o débito. Para probar esto se toma una muestra de 500 personas de las que 420 cuentan con algún tipo de tarjeta. ¿Puede aceptarse la afirmación de este banco, con un nivel de significación de 3 por ciento?

11.7 Prueba para la diferencia entre 2 proporciones con z y con χ^2

En esta sección se muestra cómo puede realizarse una prueba de hipótesis sobre la diferencia entre las proporciones de 2 poblaciones mediante 2 métodos equivalentes que, por supuesto, conducen a la misma conclusión. Los estadísticos de prueba son nuevamente la z de la distribución normal y la ji cuadrada.

11.7.1 Prueba para la diferencia entre 2 proporciones con z

El siguiente ejemplo fue resuelto en el capítulo 10, que se ocupa de las pruebas de hipótesis sobre 2 poblaciones e ilustra el procedimiento para probar la diferencia entre 2 proporciones utilizando la z normal.

Ejemplo 11.5

En un proceso de producción se encontraron 35 artículos defectuosos en una muestra aleatoria de 500 y se encontraron 20 defectuosos en otra muestra de 400 artículos provenientes de otro proceso similar que se lleva a cabo en otra fábrica. Pruebe la hipótesis de que los dos procesos producen la misma proporción de artículos defectuosos, con un nivel de significación de uno por ciento?

Solución: Las hipótesis:

$$H_0: \pi_1 = \pi_2$$
$$H_1: \pi_1 \neq \pi_2$$

Al tratarse de muestras grandes, se puede utilizar la z; con una prueba de 2 extremos y un nivel de significación de 0.01 se tiene que el valor crítico de z es 2.575, ya que $P(-2.575 \leq z \geq 2.575) = 0.01$.

La p combinada es:

$$p = \frac{x_1 + x_2}{n_1 + n_2} + \frac{35 + 20}{500 + 400} = 0.061$$

La z calculada con los datos muestrales:

$$z = \frac{p_1 - p_2}{s_{(p_1 - p_2)}} = \frac{\frac{35}{500} - \frac{20}{400}}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$= \frac{0.07 - 0.05}{\sqrt{0.061(0.939)\left(\frac{1}{500} + \frac{1}{400}\right)}} = \frac{0.02}{0.016} = 1.25$$

Debido a que la z calculada es mayor que -2.575 y menor que 2.575, no se puede rechazar la hipótesis nula y se concluye que los 2 procesos producen la misma proporción de artículos defectuosos.

11.7.2 Prueba para la diferencia entre 2 proporciones con χ^2

A continuación se contrasta la solución del ejemplo 11.5, que se resolvió utilizando la distribución normal, con su resolución utilizando la ji cuadrada.

Ejemplo 11.6

De vuelta al ejemplo anterior que explica 2 procesos de producción: uno en el que se encontraron 35 artículos defectuosos dentro de una muestra aleatoria de 500; el segundo proceso que generó 20 artículos defectuosos en otra muestra de 400. Pruebe la hipótesis de que los 2 procesos arrojan la misma proporción de artículos defectuosos, con un nivel de significación de 1% utilizando como estadístico de prueba la ji cuadrada, χ^2.

Solución: Las hipótesis son las mismas:

$$H_0: \pi_1 = \pi_2$$
$$H_1: \pi_1 \neq \pi_2$$

Se coloca en una tabla el número de artículos defectuosos y no defectuosos resultado de ambos procesos; así se resumen las frecuencias observadas:

	Proceso 1	Proceso 2	Totales
	f_o	f_o	
Defectuosos	35	20	55
No defectuosos	465	380	845
Totales	500	400	900

Para determinar las frecuencias esperadas se utilizan los totales de renglón y de columna. Por ejemplo, si se tienen 55 artículos defectuosos de un total de 900 artículos se consideraría que 55/900 = 0.0611 o 6.11% de todos los artículos que fueran defectuosos. Por ello, si en el proceso 1 se tiene un total de 500 artículos producidos, se especularía que 0.0611 × 500 = 30.55 de ellos fueran defectuosos. Estas frecuencias esperadas se anotan en el cuadro siguiente, donde también se reproducen las frecuencias observadas.

Aquí es importante notar que esas frecuencias esperadas se obtuvieron dividiendo el total de renglón, 55, entre la totalidad de los artículos, 900, con lo que se obtuvo 6.11%. Luego ese 0.0611 se multiplica por el total de la columna del proceso 1; sin embargo, se puede llegar al mismo resultado dividiendo el total de columna, 500, entre el total absoluto de 900, así resulta 500/900 = 0.556 y luego multiplicado por el total del primer renglón, 55, se obtiene la misma cantidad de frecuencia esperada: 0.556(55) = 30.58 (la pequeña diferencia se debe al redondeo). Siguiendo estos mismos razonamientos se completó la tabla 11.2 que contiene las frecuencias observadas y las esperadas.

Tabla 11.2 Frecuencias observadas y esperadas para el ejemplo 11.6

	Proceso 1		Proceso 2		Totales
	f_o	f_e	f_o	f_e	
Defectuosos	35	30.55	20	24.45	55
No defectuosos	465	469.45	380	375.55	845
Totales	500	500	400	400	900

En este caso los grados de libertad son: $gl = (c - 1)(r - 1) = (2 - 1)(2 - 1) = 1$.

Entonces el valor crítico del estadístico de prueba es:

$$P(\chi^2 \geq 3.84 \mid gl = 1) = 0.05$$

Ahora se determina el valor calculado de χ^2:

f_o	f_e	$f_o - f_e$	$(f_o - f_e)^2$	$\frac{(f_o - f_e)^2}{f_e}$
35	30.55	4.45	19.8025	0.64819967
465	469.45	–4.45	19.8025	0.04218234
20	24.45	–4.45	19.8025	0.8099182
380	375.55	4.45	19.8025	0.05272933
				1.55302954

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} = 1.55$$

Debido a que el valor calculado de χ^2, 1.55, es menor que el valor crítico, 3.84, no puede rechazarse la hipótesis nula. La conclusión, tal como se vio anteriormente, es que los 2 procesos producen la misma proporción de artículos defectuosos.

EJERCICIOS 11.7 Prueba para la diferencia entre 2 proporciones con z y con χ^2

1. Una empresa realiza evaluaciones a sus 2 productos más vendidos. Con una muestra aleatoria de 250 artículos del producto A y 200 del B, obtiene que 230 y 190 respectivamente pasan las pruebas de acuerdo con las normas de producción. El jefe de operaciones mantiene la hipótesis de que las 2 líneas de producción tienen la misma proporción de artículos que cumplen con las normas. Compruébelo con un nivel de significación de 5 por ciento.
2. El área de ventas quiere probar que la proporción de pedidos entregados a tiempo es igual entre el periodo presente y el anterior. Con una muestra aleatoria de 100 pedidos, correspondientes al periodo previo y 100 al actual, se encuentra que 80 y 92, respectivamente, fueron entregados a tiempo. Compruebe, con un nivel de significación de 2.5%, la hipótesis de que estas proporciones son iguales.
3. El departamento de ventas de una fábrica de piezas metálicas toma una muestra aleatoria de 200 piezas vendidas en la tienda 1 y otra de 100 de la tienda 2; de dichas piezas sólo 8 y 5, respectivamente, son compradas para uso doméstico. Compruebe, con un nivel de significación de 0.5%, si la proporción de piezas para uso doméstico vendidas entre las 2 tiendas es la misma.
4. Una empresa de artículos escolares supone que el porcentaje de plumas negras y rojas que duran más de 40 días es el mismo, por lo que toma una muestra aleatoria de 450 plumas negras y 300 rojas, y encuentra que 400 y 250, respectivamente, tienen el mismo tiempo de duración. Compruebe si efectivamente la proporción de plumas que rebasa la duración de 40 días es la misma para los 2 tipos. Considere un nivel de significación de 1 por ciento.
5. Un despacho de contadores toma una muestra aleatoria de 50 expedientes de clientes cuyos casos están relacionados con impuestos y 60 con contabilidades. Se descubre que sólo 6 y 8 de los expedientes, respectivamente, señalan quejas o inconformidades con el servicio prestado por el despacho. Con un nivel de significación de 1%, determine si es posible afirmar que estas proporciones son las mismas.

11.8 Relación entre las pruebas de hipótesis para proporciones con z y con χ^2

La característica principal de la relación entre estas 2 pruebas (de z y de χ^2 para probar la diferencia entre una o 2 proporciones) es que conducen a la misma conclusión. Además, existe una relación numérica entre los valores calculados correspondientes a esos 2 estadísticos de prueba: el valor calculado de χ^2 es el cuadrado del de la z. Tal como se explicó en la sección 11.2, la distribución χ^2 puede derivarse del cuadrado de la z:

$$z^2 = \frac{(X - \mu)^2}{\sigma^2}$$

No resulta extraño que el valor calculado de la χ^2 sea, precisamente, el cuadrado de la z.

En la tabla 11.3 se sintetizan los resultados obtenidos en estos últimos 4 ejemplos.

Tabla 11.3 Valores calculados de z y de χ^2 para los ejemplos 11.3 a 11.6

Ejemplo 11.3 Prueba sobre una proporción con z	1.54	Ejemplo 11.5 Prueba sobre la diferencia entre dos proporciones con z	1.25
Ejemplo 11.4 Prueba sobre una proporción con χ^2	2.38	Ejemplo 11.6 Prueba sobre la diferencia entre dos proporciones con χ^2	1.55

En la tabla se aprecia que el valor calculado de z para la prueba con una proporción fue de 1.54, en tanto que el valor calculado de χ^2 para el mismo ejercicio (ejemplo 11.4) fue de 2.38. Es posible verificar fácilmente que 2.38 es aproximadamente el cuadrado de 1.54, la pequeña diferencia se debe a errores de redondeo.

De la misma manera, el valor calculado de z para la prueba de la diferencia entre 2 proporciones fue $z = 1.25$ y el valor calculado de ji cuadrada fue $\chi^2 = 1.55$; es fácil comprobar que $1.25^2 = 1.55$ (salvo una pequeña diferencia consecuencia del redondeo).

11.9 Prueba para la diferencia entre *n* proporciones

El procedimiento anterior que se utilizó para una y para 2 proporciones puede extenderse a cualquier número, n, de ellas.

Ejemplo 11.7

En una facultad universitaria se presenta una propuesta para cambiar el plan de estudios y se desea saber si los estudiantes de cada grado tienen la misma opinión sobre la propuesta. Para ello, se toman muestras aleatorias de 100 estudiantes de cada 1 de los 4 niveles que se cursan. Los resultados se presentan en la tabla 11.4.

Tabla 11.4 Datos para el ejemplo 11.7

Nivel de los estudiantes	Tamaño de la muestra, n	Opinión sobre el cambio en el plan de estudios	
		A favor	En contra
1	100	10	90
2	100	15	85
3	100	20	80
4	100	12	88

Compruebe la hipótesis de que las proporciones de estudiantes que están a favor de modificar el plan de estudios son las mismas en todos los niveles escolares, con un nivel de significación de uno por ciento.

Solución: En este caso, las hipótesis son:

$$H_0: \pi_1 = \pi_2 = \pi_3 = \pi_4$$

H_1: Cuando menos una de las igualdades anteriores no se cumple.

El valor crítico del estadístico de prueba χ^2 es 11.34, ya que se tienen 3 grados de libertad. Estos grados de libertad son iguales al número de categorías menos 1 (4 – 1 = 3); lo mismo es decir el número de proporciones que se prueban menos 1: (4 – 1 = 3); también es igual al número de renglones menos 1 multiplicado por el número de columnas menos 1: $(r - 1)(c - 1) = (4 - 1)(2 - 1) = 3 \times 1 = 3$. En símbolos:

$$P(\chi^2 \geq 11.34 \mid gl = 3) = 0.01$$

En el siguiente cuadro se resumen las frecuencias anteriormente observadas, junto con las frecuencias esperadas y los cálculos de $\frac{(f_0 - f_e)^2}{f_e}$:

		f_o	f_o	f_e	f_e	A favor	En contra
	Totales	A favor	En contra	A favor	En contra	$\frac{(f_0 - f_e)^2}{f_e}$	$\frac{(f_0 - f_e)^2}{f_e}$
	100	10	90	14.25	85.75	1.27	0.21
	100	15	85	14.25	85.75	0.04	0.01
	100	20	80	14.25	85.75	2.32	0.39
	100	12	88	14.25	85.75	0.36	0.06
Totales	400	57	343	57	343		

Las frecuencias esperadas se calcularon dividiendo el total de renglón (100) entre el total global (400) y multiplicando este cociente por el total de la columna. También puede resolverse a la inversa y se conducirá al mismo resultado, dividiendo el total de columna (57 por ejemplo) entre el total global (400) y multiplicando este cociente por el total de renglón.

De la tabla anterior puede determinarse fácilmente el valor calculado de la ji cuadrada:

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} = 4.64$$

Este valor calculado es menor que el valor crítico de 11.34, por lo tanto no es posible rechazar la hipótesis nula y se concluye que las proporciones de los estudiantes de los diferentes niveles que están a favor del cambio en el plan de estudios son iguales.

ejercicios 11.9 Prueba para la diferencia entre *n* proporciones

1. En un estudio de mercado se preguntó a muestras independientes de 120 hombres, 100 mujeres y 100 niños si les agradaba o no el sabor de una nueva pasta dental en proceso de desarrollo; los resultados se muestran a continuación:

	Hombres	Mujeres	Niños
Les gusta	72	51	55
No les gusta	48	49	45
	120	100	100

Compruebe la hipótesis de que las proporciones de hombres, mujeres y niños a los que les gusta la pasta dental son iguales, con un nivel de significación de 0.01.

2. Se preguntó a 500 inversionistas de 4 regiones del país si consideraban que el índice del mercado accionario bajaría drásticamente en el mes siguiente. Las respuestas de cada región que estimaban que sí son: 95, 107, 100 y 83. Compruebe si las diferencias entre las proporciones correspondientes son estadísticamente significativas con un nivel de significación de 0.05.

3. Una empresa distribuidora de artículos de consumo perecederos está interesada en saber si la renovación de sus productos en los anaqueles se lleva a cabo con la frecuencia adecuada. Para saberlo, toma muestras de expendios ubicados en sus 4 regiones de distribución y obtienen los resultados que se muestran en la tabla siguiente:

Región/renovación	Norte	Pacífico	Golfo	Sureste
Renovación adecuada	25	50	53	54
Renovación no adecuada	37	52	79	148

Utilice un nivel de significación de 1% para probar la hipótesis de que son iguales las proporciones de los expendios en las 4 regiones donde la renovación de los artículos es adecuada.

4. Una universidad desea probar la hipótesis de que son iguales las proporciones de estudiantes de varias licenciaturas que tienen acceso a internet en sus casas. Para realizar la prueba, se obtienen muestras de estudiantes a quienes se les preguntó si tienen acceso a internet en su vivienda. Los resultados son los siguientes:

Licenciatura/ acceso	Administración	Filosofía	Economía	Ingeniería	Medicina
Tiene acceso	28	13	26	35	25
No tiene acceso	7	10	10	15	11

Compruebe la hipótesis con un nivel de significación de 0.05.

5. Una agencia de medios desea determinar si existen diferencias en las proporciones de personas que recuerdan un anuncio de servicios financieros publicitados en 3 medios diferentes: televisión, radio y prensa. Los resultados de un estudio de mercado son los siguientes:

	Televisión	Radio	Prensa
Personas que recordaron	30	15	10
Personas que no recordaron	80	100	112

Compruebe la hipótesis de que son iguales las 3 proporciones de personas que recordaron el anuncio, con un nivel de significación de $\alpha = 0.05$.

11.10 Pruebas de bondad de ajuste a distribuciones teóricas

Bondad de ajuste. Prueba qué tan bien se ajustan los datos observados a determinada distribución teórica.

Se revisarán ejemplos en los que se utiliza la χ^2 para probar hipótesis sobre lo que se conoce como **bondad de ajuste**. Esto quiere decir que se prueba qué tan bien se ajustan los datos observados a determinada distribución teórica.

Por ejemplo, se puede probar si determinados datos se ajustan, qué tan bien lo hacen, a una distribución normal o a una distribución Poisson. De hecho, los ejem-

plos que se presentarán en las subsecciones siguientes ilustran estos 2 casos, además de las pruebas de bondad de ajuste a una distribución binomial.

Es importante resaltar que el número de grados de libertad para este tipo de pruebas se calcula como $k - m - 1$, en donde k es el número de categorías o clases y m es el número de parámetros poblacionales que se estiman al realizar la prueba. En los ejemplos siguientes se clarificará esto.

11.10.1 Pruebas de bondad de ajuste a una distribución normal

Para realizar una prueba de este tipo se comparan las frecuencias observadas, que pudieran ajustarse a una distribución normal, con las frecuencias esperadas, determinadas a partir de la distribución normal estándar. Se ilustra el procedimiento en el siguiente ejemplo.

Ejemplo 11.8

Los datos siguientes son los montos de 220 facturas tomadas al azar. Con un nivel de significación de 0.05, ¿esos montos se ajustan a una distribución normal?

Monto de las facturas, X_i	Frecuencias observadas, f_o
0 a menos de 100	5
100 a menos de 200	7
200 a menos de 300	23
300 a menos de 400	45
400 a menos de 500	41
500 a menos de 600	43
600 a menos de 700	18
700 a menos de 800	12
800 a menos de 900	11
900 a menos de 1 000	5
Totales	210

Solución: Las hipótesis para esta prueba son:

H_0: La distribución de los montos de las facturas se ajusta a una distribución normal.

H_1: La distribución de los montos de las facturas no se ajusta a una distribución normal.

Para comenzar, en la tabla 11.5 se resumen los cálculos para determinar la media y la desviación estándar de los datos muestrales. Así, la media es:

$$\bar{X} = \frac{99\,700}{210} = 474.76$$

Y la desviación estándar:

$$s = \sqrt{\frac{7\,911\,238.10}{210}} = 194.09$$

Estimados estos 2 parámetros, se utiliza la z de la distribución normal estándar para calcular las frecuencias esperadas, la que tiene media cero desviación estándar de 1, de la siguiente manera:

$$z = \frac{X_i - \bar{X}}{s}$$

Tabla 11.5 Cálculo de la media y la desviación estándar para el ejemplo 11.8

X_i	f_o	Punto medio de clase, pm	$f_o \cdot$ pm	$(X_i - \bar{X})^2$	$f(X_i - \bar{X})^2$
0 a menos de 100	5	50	250	180 422.7	902 113.38
100 a menos de 200	7	150	1 050	105 470.3	738 292.06
200 a menos de 300	23	250	5 750	50 517.91	1 161 912.02
300 a menos de 400	45	350	15 750	15 565.53	700 448.98
400 a menos de 500	41	450	18 450	613.1519	25 139.23
500 a menos de 600	43	550	23 650	5 660.771	243 413.15
600 a menos de 700	18	650	11 700	30 708.39	552 751.02
700 a menos de 800	12	750	9 000	75 756.01	909 072.11
800 a menos de 900	11	850	9 350	140 803.6	1 548 839.91
900 a menos de 1 000	5	950	4 750	225 851.2	1 129 256.23
	210		99 700		7 911 238.10

Luego se estandarizan los límites de los intervalos en los que están divididas las frecuencias. Por ejemplo, el extremo izquierdo, es decir, la clase de 0 a menos de 100 está limitada del lado derecho por ese valor (100) que en términos de la z, es igual a:

$$z = \frac{X_i - \bar{X}}{s} = \frac{100 - 474.76}{194.09} = \frac{-374.76}{194.09} = -1.93$$

El área bajo la curva y a la izquierda de este valor es $P(-\infty \leq z \leq -1.93) = 0.0268$, que en términos de las unidades originales, el monto de las facturas, es $P(0 \leq X \leq 100) = 0.0268$. Ahora, para el segundo intervalo, el que va de 100 a 200:

$$z = \frac{X_i - \bar{X}}{s} = \frac{200 - 474.76}{194.09} = \frac{-274.76}{194.09} = -1.42$$

El correspondiente valor de área o probabilidad es $P(-1.93 \leq z \leq -1.42) = 0.051$, que en términos de las unidades originales, el monto de las facturas, es $P(100 \leq X \leq 200) = 0.051$.

Si se calculan, de la misma manera, las probabilidades de todos los intervalos se obtienen los valores de la tabla 11.6.

Tabla 11.6 Valores de z y probabilidades correspondientes a los 10 intervalos de datos del ejemplo 11.8

Límite de intervalo	z	Área	Frecuencias esperadas	Frecuencias esperadas agrupadas
100	–1.93	0.0268	5.63	5.63
200	–1.42	0.051	10.71	10.71
300	–0.90	0.1063	22.32	22.32
400	–0.39	0.1642	34.48	34.48
500	0.13	0.2034	42.71	42.71
600	0.65	0.1905	40.01	40.01
700	1.16	0.1348	28.39	28.39
800	1.68	0.0765	16.07	16.07
900	2.19	0.0322	6.76	9.76
1 000		0.0143	3.00	

La determinación del área correspondiente al intervalo 400 a 500 es la única que requiere cálculos adicionales, ya que la media cae dentro de este intervalo, como puede verse en la figura 11.3. Así que el área correspondiente debe calcularse en 2 pasos: de 400 hasta la media, 474.76, y de esta media hasta 500. Se llega al área anotada al realizar las operaciones correspondientes y sumando ambas probabilidades.

En la tabla 11.6 también se calcularon las frecuencias esperadas simplemente multiplicando la probabilidad correspondiente a cada intervalo por 210, el número de elementos de la muestra. Adicionalmente, se agruparon en la columna del extremo derecho las frecuencias esperadas para los 2 últimos intervalos, 800-900 y 900-1 000, porque las frecuencias esperadas del último intervalo son 3.00 y esta prueba de ji cuadrada requiere que haya cuando menos 5 frecuencias teóricas (esperadas) en cada celda. Si los datos en los que se basa la prueba generan celdas con frecuencias esperadas menores a 5, se requiere combinar celdas para asegurar que todas contengan al menos 5 esperadas.

Sumadas las 2 últimas frecuencias esperadas, 6.76 y 3.00, resulta 9.76 de las frecuencias esperadas de este intervalo combinado. En la figura 11.3 se muestra la gráfica de estas frecuencias esperadas agrupadas, se observa que tienen forma aproximadamente normal.

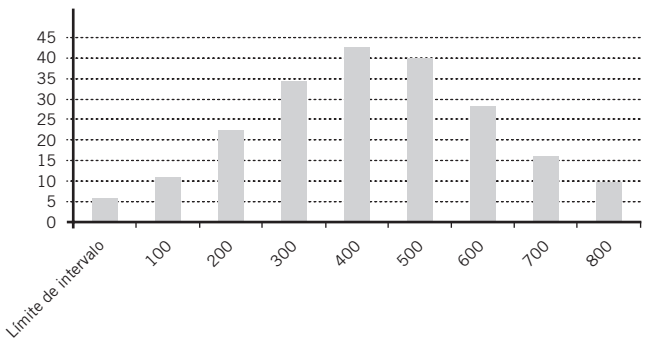


Figura 11.3 Frecuencias esperadas agrupadas correspondientes a los 10 intervalos de datos del ejemplo 11.8.

En la tabla 11.7 se reúnen las frecuencias observadas y las esperadas; se realizan las operaciones necesarias para calcular la ji cuadrada.

Tabla 11.7 Cálculos para obtener la χ^2 en el ejemplo 11.8

X_i	f_o	Agrupadas	f_e	$\frac{(f_o - f_e)^2}{f_e}$
0 a menos de 100	5	5	5.63	0.0705
100 a menos de 200	7	7	10.71	1.2852
200 a menos de 300	23	23	22.32	0.0207
300 a menos de 400	45	45	34.48	3.2097
400 a menos de 500	41	41	42.71	0.0685
500 a menos de 600	43	43	40.01	0.2234
600 a menos de 700	18	18	28.39	3.8025
700 a menos de 800	12	12	16.07	1.0308
800 a menos de 900	11	16	9.76	3.9895
900 a menos de 1 000	5			
Totales	210			13.7008

Finalmente quedaron 9 categorías (porque se combinaron las 2 últimas) y se estimaron 2 parámetros poblaciones (la media y la desviación estándar); los grados de libertad son $k - m - 1 = 9 - 2 - 1 = 6$, por lo que el valor crítico de la χ^2 para un nivel de significación de 0.05 es 12.592:

$$P(\chi^2 \geq 12.592 \mid gl = 6) = 0.05$$

El valor calculado de la χ^2, 13.7, es mayor que este valor crítico, así que se rechaza la hipótesis nula y se concluye que los montos de esa muestra de facturas no se ajustan a una distribución normal.

11.10.2 Pruebas de bondad de ajuste a una distribución Poisson

Para realizar una prueba de este tipo se comparan las frecuencias observadas, que pudieran ajustarse a una distribución Poisson, con las frecuencias esperadas, determinadas, precisamente, a partir de la distribución Poisson, tal como en el ejemplo anterior. Estas frecuencias esperadas se determinan fácilmente utilizando la función Poisson de Excel, como se verá a continuación.

Ejemplo 11.9

Se desea probar, con un nivel de significación del 0.05, si el número de descomposturas de máquinas, por hora en una línea de ensamble, sigue una distribución de Poisson. Para probarlo se toma una muestra aleatoria de 50 horas; las cantidades de descomposturas obtenidas se muestran en las 2 primeras columnas de la tabla siguiente:

Descomposturas por hora, X	Frecuencias observadas, f_o	$X \cdot f_o$
0	0	0
1	5	5
2	9	18
3	11	33
4	8	32
5	4	20
6	3	18
7	2	14
8	1	8
9	0	0
10	1	10
	44	158

Solución: Las hipótesis para esta prueba son:

H_0: La distribución de descomposturas de máquinas por hora se ajusta a una distribución Poisson.
H_1: La distribución de descomposturas de máquinas por hora no se ajusta a una distribución Poisson.

En la misma tabla se obtuvieron los productos de las descomposturas por hora, multiplicadas por su frecuencia, $X \cdot f_o$, y se determinó la media de esta muestra:

$$\bar{X} = \lambda = \frac{158}{44} - 3.5909$$

Con esta media puede utilizarse la función Poisson de Excel para determinar las probabilidades teóricas. Por ejemplo, con la función =POISSON(0,3.5909,) se obtiene el valor 0.0276, probabilidad de cero descomposturas, en una distribución Poisson con una media de 3.5909. Con esta función se generó la distribución de frecuencias Poisson de la tabla 11.8, donde también se incluyeron las frecuencias observadas y las esperadas agrupadas para respetar la **regla de las 5**, que obliga a utilizar al menos 5 frecuencias esperadas en cada categoría.

Regla de las 5. Obliga a utilizar al menos 5 frecuencias esperadas en cada categoría.

Tabla 11.8 Frecuencias observadas y esperadas para el ejemplo de prueba de bondad de ajuste a una distribución Poisson

Descomposturas por hora, X	Frecuencias observadas, f_o	Probabilidades Poisson, P	Frecuencias esperadas, P·44	Frecuencias esperadas agrupadas
0	0	0.0276	1.2	
1	5	0.0990	4.4	5.6
2	9	0.1778	7.8	7.8
3	11	0.2128	9.4	9.4
4	8	0.1910	8.4	8.4
5	4	0.1372	6.0	6.0
6	3	0.0821	3.6	6.8
7	2	0.0421	1.9	
8	1	0.0189	0.8	
9	0	0.0075	0.3	
10	1	0.0027	0.1	
11	0	0.0009	0.0	
12	0	0.0003	0.0	
13	0	0.0001	0.0	
14	0	0.0000	0.0	
	44	1.0000		

Repasando, en este caso las frecuencias esperadas se obtienen multiplicando las probabilidades Poisson por 44, el número de horas muestreadas. Finalmente, la última columna contiene solamente 6 frecuencias esperadas, ya que se agruparon en una sola categoría los 2 primeros renglones (0 y 1 descomposturas) y los últimos 9 (de 6 a 14 descomposturas) para cumplir el requerimiento de la *regla del 5*.

Con $n = 6$ categorías y habiendo estimado la media de la distribución con los datos de la muestra, los grados de libertad son $k - m - 1 = 6 - 1 - 1 = 4$. Tomando en cuenta los grados de libertad, el valor crítico del estadístico de prueba es:

$$P(\chi^2 \geq 9.49 \mid gl = 4) = 0.05$$

Se resumen los cálculos de la χ^2:

f_o	f_e	$\frac{(f_o - f_e)^2}{f_e}$
5	5.57	0.0583
9	7.82	0.1774
11	9.36	0.2863
8	8.41	0.0195
4	6.04	0.6870
3	6.80	2.1263
		3.3549

De manera que, como el valor calculado de χ^2 (3.35) es menor que el valor crítico (9.49) no se rechaza la hipótesis nula y se concluye que la distribución de las descomposturas de máquinas por hora se ajusta a una distribución Poisson.

11.10.3 Pruebas de bondad de ajuste a una distribución binomial

Para llevar a cabo esta prueba se comparan las frecuencias observadas, que pudieran ajustarse a una distribución binomial, con las frecuencias esperadas, determinadas, precisamente a partir de la distribución binomial.

EJEMPLO 11.10

Un gerente de operaciones está interesado en construir un modelo matemático que describa el comportamiento de las descomposturas de las máquinas utilizadas en la producción. Considera que ese comportamiento podría describirse mediante una distribución binomial. Para evaluar esta posibilidad toma una muestra de 30 semanas y cuenta las descomposturas en cada máquina. Los resultados son los siguientes:

	Núm. de descomposturas por semana						
Núm. de semanas	0	1	2	3	4	5	6 o más
	7	10	10	1	1	1	0

Solución: Las hipótesis para esta prueba son:

H_0: La distribución del número de descomposturas por semana se ajusta a una distribución binomial.
H_1: La distribución del número de descomposturas por semana se ajusta a una distribución binomial.

Con los datos observados se calcula la probabilidad de que una máquina se descomponga en una semana cualquiera. El cálculo:

$$p = \frac{\text{total de máquinas que se descomponen}}{(\text{total de máquinas})(\text{total de semanas})}$$

$$p = \frac{0(7) + 1(10) + 2(10) + 3(1) + 4(1) + 5(1) + 6(0)}{10(30)}$$

$$= \frac{7 + 10 + 20 + 3 + 4 + 5}{300} = 0.163$$

Con este valor de probabilidad para la descompostura de una máquina (estimado a partir de los datos muestrales) se construye la distribución de probabilidad binomial teórica a partir de la función de probabilidad estudiada en el capítulo 5 (lo cual también puede llevarse a cabo con la función "Distr.Binom" de Excel):

$$P(x) = C_n^n p^X q^{n-x} = \frac{n!}{x!(n-x)!} p^X q^{n-x}$$

$P(X) = C_x^{10}\left(0.163^x\right)\left(0.837^{10-x}\right)$	$30 \cdot P(X)$
$P(X = 0) = 0.1688$	5.06
$P(X = 1) = 0.3286$	9.86
$P(X = 2) = 0.2880$	8.64
$P(X = 3) = 0.1496$	4.49
$P(X = 4) = 0.0510$	1.53
$P(X = 5) = 0.0119$	0.36
$P(X \geq 6) = 0.0022$	0.06

Como los valores de las frecuencias esperadas deben ser de al menos 5, se suman los últimos 4 renglones:

$P(X) = C_x^{10}\left(0.163^x\right)\left(0.837^{10-x}\right)$	$30 \cdot P(X)$
$P(X = 0) = 0.1688$	5.06
$P(X = 1) = 0.3286$	9.86
$P(X = 2) = 0.2880$	8.64
$P(X \geq 3) = 0.2146$	6.44

De acuerdo con la distribución binomial, ya se tienen las frecuencias observadas (las de la muestra) y las esperadas (las teóricas). En el cuadro siguiente se resumen estos datos, también se combinaron los datos de 3 o más máquinas descompuestas para los valores observados en la muestra, por el requerimiento de la regla de las 5. Asimismo se anotan los cálculos necesarios para determinar el valor de la χ^2.

f_o	f_e	$f_o - f_e$	$(f_o - f_e)^2$	$\frac{(f_o - f_e)^2}{f_e}$
7	5.06	1.94	3.7636	0.7438
10	9.86	0.14	0.0196	0.0020
10	8.64	1.36	1.8496	0.2141
3	6.44	–3.44	11.8336	1.8375
			Suma	2.7974

En este caso se estimó la p a partir de los datos muestrales por lo que se tiene $m = 1$, y los grados de libertad son: $gl = k - m - 1 = 4 - 1 - 1 = 2$.

El valor crítico del estadístico de prueba es:

$$P\left(\chi^2 \geq 5.99 \mid gl = 2\right) = 0.05$$

El valor calculado de la χ^2, 2.7974, es menor que su valor crítico, 5.99, entonces no se rechaza H_0 y se concluye que la distribución de las descomposturas de esas máquinas sí se ajusta adecuadamente a una distribución binomial.

EJERCICIOS 11.10 Pruebas de bondad de ajuste a distribuciones teóricas

Pruebas de bondad de ajuste a una distribución normal

1. Una compañía de servicios de mensajería registra el peso de 70 paquetes elegidos al azar. Determine si el peso de los paquetes se ajusta a una distribución normal con un nivel de significación de 0.01.

Peso de los paquetes (en g), X_i	f_o
0 a menos de 50	
50 a menos de 100	2
100 a menos de 150	3
150 a menos de 200	8
200 a menos de 250	15
250 a menos de 300	15
300 a menos de 350	15
350 a menos de 400	10
400 a menos de 450	1
450 a menos de 500	1
Totales	70

2. Un ingeniero quiere evaluar la calidad de las computadoras de una compañía, toma una muestra aleatoria de 63 de equipos y quiere saber si la calidad de los mismos se ajusta a una distribución normal con un nivel de significación de 5 por ciento.

Calificación de la evaluación de calidad, X_i	f_o
0 a menos de 1	1
1 a menos de 2	7
2 a menos de 3	15
3 a menos de 4	25
4 a menos de 5	15
Totales	63

3. El área de recursos humanos de una organización quiere averiguar si las horas diarias que trabajan los empleados se ajustan a una distribución normal con un nivel de significación de 2.5%. Toma una muestra aleatoria de 300 empleados y obtiene los siguientes datos:

Horas de trabajo diario, X_i	f_o
0 a menos de 2	40
2 a menos de 4	45
4 a menos de 6	100
6 a menos de 8	115
Totales	300

4. Al personal docente de una universidad se le aplicó una prueba de conocimientos que califica de 0 a 100. Determine si las calificaciones se distribuyen de forma normal, considerando un nivel de significación de 0.5, de acuerdo con los siguientes resultados de la muestra aleatoria de 100 profesores.

Calificación, X_i	f_o
0 a menos de 10	1
10 a menos de 20	3
20 a menos de 30	3
30 a menos de 40	6
40 a menos de 50	5
50 a menos de 60	17
60 a menos de 70	22
70 a menos de 80	30
80 a menos de 90	8
90 a menos de 100	5
Totales	100

5. El administrador de un negocio quiere probar, con un nivel de significación de 0.5%, si el importe total de ventas diarias se ajusta a una distribución normal. Con una muestra aleatoria de 125 días se obtienen los siguientes resultados:

Importe total de ventas diarias en pesos, X_i	f_o
0 a menos de 3 000	35
3 000 a menos de 6 000	45
6 000 a menos de 9 000	20
9 000 a menos de 12 000	20
12 000 a menos de 15 000	5
Totales	125

Pruebas de bondad de ajuste a una distribución Poisson

6. Con un nivel de significación de 0.03 se quiere saber si el número de clientes que llega a un supermercado durante un intervalo de 5 min sigue una distribución de Poisson. Para saberlo se toma una muestra aleatoria de 30 intervalos. Éstos son los resultados:

Clientes por intervalo, X	Frecuencias observadas, f_o	$X \cdot f_o$
0	0	0
1	1	1
2	0	0
3	6	9
4	5	20
5	4	20
6	2	12
7	3	42
8	6	48
9	2	18
10 o más	1	10
	30	180

7. Se cree que el número de accidentes automovilísticos que ocurren a diario en determinada ciudad sigue una distribución Poisson. Para probarlo se toma una muestra aleatoria de 50 días del año pasado y se obtienen los siguientes resultados:

Núm. de accidentes, X	Frecuencias observadas, f_o	$X \cdot f_o$
0	0	0
1	2	2
2	5	10
3	7	21
4	5	20
5	6	30
6	7	42
7	4	28
8	2	16
9	5	45
10 o más	7	70
	50	284

Pruebe esa hipótesis con un nivel de significación de 0.05.

8. Se desea saber si el número de llamadas telefónicas que entran al conmutador de una empresa, durante intervalos de 1 minuto, tiene una distribución de Poisson. Se toma una muestra de 100 llamadas con un nivel de significación de 0.10, se obtiene la siguiente tabla:

Núm. de llamadas por intervalo, X	Frecuencias observadas, f_o	$X \cdot f_o$
0	2	0
1	3	3
2	4	8
3	6	18
4	5	20
5	5	25
6	7	42
7	6	42
8	7	56
9	9	81
10	8	80
11	5	55
12	6	72

Núm. de llamadas por intervalo, X	Frecuencias observadas, f_o	$X \cdot f_o$
13	8	104
14	9	126
15 o más	10	150
	100	882

9. Se quiere probar si los pedidos que registra un vendedor vía telefónica sigue una distribución de Poisson. Se toma una muestra de 60 días hábiles de los meses recientes, y se obtienen los siguientes datos:

Núm. de pedidos levantados, X	Frecuencias observadas, f_o	$X \cdot f_o$
0	5	0
1	10	10
2	12	24
3	15	45
4	9	36
5	8	40
6 o más	1	6
	60	161

Realice la prueba con un nivel de significación de 0.025.

Pruebas de bondad de ajuste a una distribución binomial

10. Diariamente la academia de matemáticas realiza evaluaciones a grupos de 10 alumnos. En una muestra aleatoria de 35 días se cuentan los exámenes reprobatorios de cada día. Se obtienen los resultados que se presentan a continuación:

Exámenes reprobados diariamente	Núm. de días
0	5
1	7
2	9
3	6
4	4
5 o más	4

Determine si el comportamiento del número de exámenes reprobatorios por día se ajusta a una distribución binomial con un nivel de significación de 0.025.

11. El jefe de una empresa de venta de refacciones automotrices considera que el robo de productos exhibidos en mostrador durante la última temporada puede describirse mediante una distribución binomial. Toma una muestra aleatoria de 7 semanas y cuenta los artículos faltantes obteniendo los siguientes resultados:

Productos faltantes	Núm. de semanas
0	2
1	4
2	5
3	5
4	3
5	1
6 o más	0

Determine, con un nivel de significación de 0.01, si estos datos se adecuan a una distribución binomial considerando que en el mostrador se exhiben 30 productos semanalmente.

12. Para evaluar la efectividad de un producto, un laboratorio evalúa diariamente a 15 pacientes que reciben el tratamiento. Obtuvo los siguientes resultados de una muestra aleatoria de 50 días:

Núm. de personas a las cuales no les funciona el producto	Núm. de días
0	8
1	14
2	15
3	7
4	5
5 o más	1

Determine con un nivel de significación de 0.005 si los datos se ajustan a una distribución binomial.

13. El supervisor de personal de una empresa manufacturera toma una muestra aleatoria de 25 salas de confección para evaluar la asistencia de los empleados. Los resultados son los siguientes:

Núm. de empleados faltantes	Núm. de salas
0	1
1	2
2	7
3	7
4	3
5	2
6	1
7	1
8	1
9 o más	0

Determine, con un nivel de significación de 0.05, si los datos presentados se ajustan a una distribución binomial; tome en cuenta que en cada sala debe haber 20 empleados.

11.11 Pruebas de bondad de ajuste entre distribuciones empíricas

En este tipo de pruebas se intenta evaluar si 2 o más poblaciones tienen distribuciones similares entre sí. Por ejemplo, conocer si la distribución de las calificaciones en cierta prueba es la misma en 2 poblaciones diferentes (ver el ejemplo 11.11); saber si los hábitos (ver televisión, navegar en internet, o cualquier otra característica) se distribuyen de la misma manera en diferentes grupos de edades; averiguar si las razones por las que los consumidores compran cierto producto se distribuyen de la misma manera en estratos sociales diferentes. Se presentan en seguida un par de ejemplos.

EJEMPLO 11.11

Se aplica una prueba de rendimiento en 2 escuelas y se toman muestras de 100 estudiantes de cada una. Los resultados se observan en la tabla 11.9. Determine si la distribución del rendimiento de los estudiantes es la misma en las 2 escuelas, con un nivel de significación de 0.01.

Tabla 11.9 Datos del ejemplo 11.11

Calificación	Escuela A	Escuela B
Menos de 6	10	10
7	20	10
8	30	40
9	20	30
10	20	10
Total	100	100

Observe que se compararán 5 categorías, los 5 niveles de calificación. Se evaluará si la distribución de las calificaciones en estas categorías es igual en las 2 escuelas; en otras palabras, se evaluará si esa distribución es homogénea en ambas instituciones educativas.

Solución: En esta prueba podría considerarse que la hipótesis nula es que si ambas muestras tienen el mismo comportamiento, entonces se diría que provienen de la misma población. También es posible plantear que ambas muestras tienen la misma distribución:

H_0: Ambas muestras tienen la misma distribución.
H_1: Las muestras no tienen la misma distribución.

Dado el nivel de significación $\alpha = 0.01$, y con 4 grados de libertad, el valor crítico de χ^2 es 13.28, dado que, $P(\chi^2 \mid gl = 4 > 13.28) = 0.01$.

El procedimiento a seguir es el mismo que para las pruebas de independencia: se obtienen las frecuencias esperadas combinando los valores observados.

En la tabla siguiente se muestran entre paréntesis las frecuencias esperadas. Las f_e se calculan dividiendo el total de cada renglón entre el total de los 200 alumnos y multiplicando este cociente por 100, el total de la columna.

Calificación	Escuela A	Escuela B	Total
Menos de 6	10 (10)	10 (10)	20
7	20 (15)	10 (15)	30
8	30 (35)	40 (35)	70
9	20 (25)	30 (25)	50
10	20 (15)	10 (15)	30
Total	100	100	200

Finalmente, en la tabla que se presenta a continuación se muestran las operaciones necesarias para determinar el valor de χ^2.

Escuela A	Escuela B	Escuela A $\frac{(f_o - f_e)^2}{f_e}$	Escuela B $\frac{(f_o - f_e)^2}{f_e}$
10 (10)	10 (10)	0	0
20 (15)	10 (15)	1.7	1.7
30 (35)	40 (35)	0.7	0.7
20 (25)	30 (25)	1.0	1.0
20 (15)	10 (15)	1.7	1.7
		5.1	5.1

Debido a que el valor calculado de $\chi^2 = 10.2$ es menor que el valor crítico de $\chi^2 = 13.28$, no es posible rechazar la hipótesis nula y se concluye que la distribución de las calificaciones de los estudiantes de las 2 escuelas es homogénea.

EJEMPLO 11.12

En una fábrica quiere evaluarse si el número de descomposturas de máquinas es igual todos los días de la semana. Para hacerlo se obtuvieron datos de descomposturas promedio por cada día de la semana. Los resultados se muestran a continuación:

Día	Máquinas descompuestas
Lunes	13
Martes	14
Miércoles	6
Jueves	10
Viernes	8
Sábado	7
Domingo	12

Pruebe la hipótesis con un nivel de significación de 0.05.

Solución: Las hipótesis:

H_0: El número de descomposturas de máquinas es igual todos los días de la semana.
H_1: El número de descomposturas de máquinas no es igual todos los días de la semana.

Se tienen 6 grados de libertad; por lo tanto, el valor crítico de la χ^2 es 12.5916, ya que:

$$P(\chi^2 \mid gl = 6 > 12.5916) = 0.05$$

Además, como la hipótesis nula plantea que se descompone el mismo número de máquinas todos los días y la suma de las descomposturas observadas es 70, el número de descomposturas debería ser de 10 diarias en caso de que esta hipótesis fuera cierta. En la tabla siguiente se resumen estos datos y se realizan las operaciones necesarias para calcular la χ^2:

Día	Máquinas descompuestas, f_o	Máquinas descompuestas según la hipótesis nula, f_e	$\frac{(f_o - f_e)^2}{f_e}$
Lunes	13	10	0.9
Martes	14	10	1.6
Miércoles	6	10	1.6
Jueves	10	10	0
Viernes	8	10	0.4
Sábado	7	10	0.9
Domingo	12	10	0.4
Totales	70	70	5.8

La χ^2 calculada es 5.8, menor que el valor crítico, 12.59; entonces no es posible rechazar la hipótesis nula y se concluye que, efectivamente, el número de máquinas descompuestas es igual en todos los días de la semana.

EJERCICIOS 11.11 Pruebas de bondad de ajuste entre distribuciones empíricas

1. Se tomó una muestra aleatoria de la terminación numérica de los billetes de lotería ganadores. Se registraron los siguientes resultados:

Dígito final	Frecuencia observada
0	8
1	13
2	9
3	8
4	16
5	13
6	10
7	7
8	15
9	11

Pruebe, con un nivel de significación de 0.01, la hipótesis de que cada dígito tiene la misma probabilidad de ser terminación de lotería ganadora.

2. En una muestra aleatoria de 200 personas a las que se les preguntó si estaban de acuerdo o no con la prohibición de fumar en lugares públicos se encontraron los siguientes resultados:

	Edad		
	18 a 35	36 a 55	Más de 55
De acuerdo con la prohibición	76	54	70

Pruebe la hipótesis de que las personas en todos los grupos de edad tienen la misma disposición en cuanto a la prohibición de fumar en lugares públicos, con un nivel de significación de 0.05.

3. En una oficina de administración de personal se desea investigar si la impuntualidad de los empleados se distribuye igual en todos los días de la semana. Para hacerlo, se toma una muestra aleatoria de reportes de retardos durante un periodo de 6 meses, y se obtienen los siguientes resultados:

Día	Núm. de retardos
Lunes	60
Martes	35
Miércoles	50
Jueves	47
Viernes	58

Pruebe la hipótesis de que el número de retardos es igual para todos los días de la semana con un nivel de significación de 0.02.

11.12 Pruebas sobre la independencia entre 2 variables

Se explicará el procedimiento a seguir para probar hipótesis sobre la independencia entre 2 variables. Es importante distinguir una característica sobre estos casos; se trata de una muestra de elementos clasificados de acuerdo con 2 variables y 2 criterios. Por ejemplo, saber si la orientación política es independiente del nivel educativo; si la calidad de cierto artículo es independiente del turno en el que se fabrica (matutino, vespertino, mixto); si el nivel de ingresos es independiente de la puntualidad o morosidad en los pagos de créditos.

No debe ignorarse que estas pruebas de independencia se realizan con datos agrupados en tablas de contingencias, como las que se utilizaron para las pruebas sobre proporciones; sin embargo, en éstas sólo había 2 renglones. En tanto que para las pruebas de independencia el número de categorías suele ser de más de 2 para ambas variables de clasificación.

En el ejemplo siguiente se ilustra la forma como se utiliza la distribución χ^2 para este tipo de pruebas.

EJEMPLO 11.13

Una empresa de investigación de mercados desea saber si la marca de ciertos automóviles depende de la zona en la que habitan sus propietarios. Para investigarlo, toma una muestra aleatoria de 600 propietarios con sus autos e identifica qué marca poseen y en qué zona de la ciudad habitan. En la tabla 11.10 se muestran los resultados. La empresa decide realizar la prueba con un nivel de significación de 0.01.

Tabla 11.10 Propietarios de autos clasificados por la marca y la zona en la que habitan

Zona de residencia	Marca de automóvil			Total
	A	B	C	
I	64	75	35	174
II	71	70	63	204
III	61	76	85	222
Total	196	221	183	600

Solución: Las hipótesis para esta prueba son:

H_0: La marca de auto que poseen los propietarios es independiente de la zona de la ciudad en la que habitan.

H_1: La marca de auto que poseen los propietarios sí depende de la zona de la ciudad en la que habitan.

Con $\alpha = 0.01$ y con 4 grados de libertad, ya que $(c - 1)(r - 1) = 2 \times 2 = 4$, el valor crítico de la χ^2 es 13.277:

$$P(\chi^2 \geq 13.277 \mid gl = 4) = 0.01$$

En la tabla 11.11 se muestran las frecuencias observadas junto con las frecuencias esperadas; éstas se calcularon dividendo el total de renglón o de columna entre el total global (600) y multiplicando este cociente por el total de columna o renglón, según se haya realizado la primera operación.

Tabla 11.11 Frecuencias observadas y esperadas; cálculos para el ejemplo 11.13

f_o			f_e			$(f_o - f_e)^2/f_e$		
64	75	35	56.84	64.09	53.07	0.90	1.86	6.15
71	70	63	66.64	75.14	62.22	0.29	0.35	0.01
61	76	85	72.52	81.77	67.71	1.83	0.41	4.42
196	221	183	196	221	183			

La suma de los cuadrados de las diferencias entre las frecuencias observadas y esperadas, divididas entre las esperadas (las 3 columnas de la derecha de la tabla) da un total de 16.21 que es el valor calculado del estadístico de prueba, χ^2. Este valor calculado es mayor que el valor crítico, 13.277, así que se rechaza la hipótesis nula y se concluye que, efectivamente, la marca de auto que poseen los propietarios de automóviles de esa ciudad sí depende de la zona en la que habitan.

EJERCICIOS 11.12 Pruebas sobre la independencia entre 2 variables

1. Se desea probar si el tipo de defecto observado en las unidades producidas en una planta manufacturera es independiente del turno en el que se fabrican. Se toma una muestra de productos de los diferentes turnos y se obtienen los resultados que se muestran en la tabla:
Pruebe la hipótesis con un nivel de significación de 0.01.

	Tipo de defecto				
Turno	A	B	C	D	Totales
Matutino	16	22	46	14	98
Vespertino	27	32	35	6	100
Nocturno	34	18	50	21	123
Totales	77	72	131	41	321

2. Para saber si la calidad de la educación primaria depende de la ubicación de la escuela, se tomó una muestra de escuelas que arrojó los siguientes resultados:

	Ubicación de la escuela			
Rendimiento escolar	Urbana	Suburbana	Rural	Totales
Excelente	20	80	20	120
Bueno	60	60	40	160
Regular	20	60	40	120
Deficiente	10	20	20	50
Totales	110	220	120	450

Pruebe la hipótesis de independencia con un nivel de significación de 0.05.

3. Una empresa de servicios de capacitación para empleados de nuevo ingreso de diversas compañías analiza la posible relación entre el desempeño de los empleados durante los cursos de capacitación y su desempeño en el trabajo. Obtiene una muestra aleatoria de empleados que ha capacitado y registra los resultados que se muestran en la tabla siguiente:

	Desempeño en la capacitación			
Rendimiento en el trabajo	Bueno	Normal	Inferior al normal	Totales
Bueno	24	61	30	115
Normal	29	80	61	170
Deficiente	10	50	64	124
Totales	63	191	155	409

Realice la prueba con un nivel de significación de 0.01.

4. Con el fin de determinar si existe relación entre el tipo de sangre y los resultados de la aplicación de un medicamento para curar la gripe, se llevó a cabo un estudio y se obtuvieron los siguientes resultados:

	Tipo de sangre			
	A	B	AB	O
Sanación total	214	198	182	189
Mejoría	52	44	38	56
Sin sanación	33	56	81	54

Con un nivel de significación de 0.05, ¿existe relación entre las 2 variables?

5. Se analizaron los limones producidos en 4 estados del país para evaluar si existe relación entre las regiones y la calidad de los mismos. Los resultados que se obtuvieron fueron:

	Colima	Veracruz	Guerrero	Michoacán
Sin defectos	880	690	1130	910
Defectos leves	250	175	305	260
No comestible	80	116	178	100

¿La calidad de los limones es independiente del estado en donde se cosechan? Realice la prueba con un nivel de significación de 0.01.

11.13 Pruebas paramétricas y pruebas no paramétricas

Los procedimientos de pruebas de hipótesis que se revisaron hasta el capítulo anterior son conocidos como *métodos paramétricos* o *métodos clásicos* y se distinguen, como se verá, sin que se haga una distinción enteramente tajante de las pruebas *no paramétricas* o *métodos no paramétricos*. La diferenciación entre unos y otros se hace contrastando las características de los métodos paramétricos:

1. Las pruebas se realizan sobre valores que se supone tienen los parámetros de una o más poblaciones (por ejemplo, una prueba sobre la media de una población, μ, o sobre la diferencia entre las proporciones de 2 poblaciones, $\pi_1 - \pi_2$). Precisamente esta característica es la que le da el nombre de *pruebas paramétricas*.
2. Los datos deben estar medidos en escala que sea cuando menos de intervalo. Esto permite, entre otras cosas, calcular medias y desviaciones estándar utilizadas para realizar pruebas de hipótesis. Conviene recordar la clasificación de las escalas de medición que se elaboró en la sección 1.6 del primer capítulo.

En el capítulo que ahora compete se revisaron pruebas tanto paramétricas como no paramétricas. Las pruebas sobre una varianza o sobre una, 2 o n proporciones se clasificarían como pruebas paramétricas, en tanto que las demás pruebas son no paramétricas porque no se ajustan a cuando menos uno de los 2 criterios.

Las pruebas de bondad de ajuste no se realizan sobre ningún parámetro poblacional; consiste en evaluar si determinados datos se ajustan o no a alguna distribución teórica o empírica. Las pruebas para la

independencia entre 2 variables también son pruebas no paramétricas porque se prueba la independencia entre variables y no algo relacionado con algún parámetro poblacional.

En el capítulo 17 "Estadística no paramétrica" se revisarán otras pruebas de hipótesis de este tipo.

11.14 Excel y la distribución ji cuadrada

En la sección 11.3 se explicó que la función de Excel Prueba.Chi.Inv arroja la probabilidad de que la χ^2 sea mayor o igual que determinado valor para un número de grados de libertad especificado. Esta información es la que se registra en las tablas de áreas bajo la curva de esta distribución, misma que permite determinar los valores críticos de la distribución para realizar pruebas de hipótesis.

Este paquete de Microsoft tiene otras 2 funciones relacionadas con esta distribución: DISTR.CHI (*x*, grados de libertad) y PRUEBA.CHI (rango de frecuencias observadas, rango de frecuencias esperadas).

11.14.1 Función Distr.Chi

La función Distr.Chi da como resultado la probabilidad de que ocurra una variable ji cuadrada de determinada magnitud, *x*, para los grados de libertad especificados. Esta función puede servir para resolver pruebas de hipótesis utilizando el método de π que ya se ha revisado.

- Se ilustra su uso con algunos de los ejemplos resueltos en este capítulo.
- En el ejemplo 11.1 se probó la hipótesis sobre si la varianza del peso de ciertos tubos no rebasaba 5.3 g^2; se encontró una χ^2 calculada a partir de los datos muestrales de 45.14 y se estableció un nivel de significación de 1% con 29 grados de libertad. En el ejemplo no se rechazó H_0 porque ese valor calculado de ji cuadrada resultó ser inferior al valor crítico de 49.588. Ahora, con la función =DISTR.CHI(45.14,29) se obtiene que la probabilidad de obtener ese valor de χ^2, 45.14, teniendo 29 grados de libertad es de 0.029, es decir, 2.9 por ciento.
- Esta probabilidad es superior al nivel de significación especificado de 0.01, al igual que en el ejemplo; entonces no se rechaza la hipótesis nula.

En el ejemplo 11.7 se intentó probar la igualdad de 4 proporciones con respecto a su opinión sobre un proyecto para cambiar el plan de estudios de una universidad. Se estableció un nivel de significación de 0.01. Una vez obtenidos los datos y realizados los cálculos, se encontró una χ^2 calculada de 4.64 con 3 grados de libertad. En el ejemplo no se rechazó la hipótesis nula porque la χ^2 calculada resultó ser menor que la crítica.

Si ahora se utiliza la función =DISTR.CHI(4.64,3) se obtiene que la probabilidad de obtener ese valor de χ^2, 4.64, con los 3 grados de libertad, es de 0.200137, es decir, de 20%. Este valor de probabilidad es muy superior al nivel de significación; entonces no se rechaza la hipótesis nula y se concluye que son iguales las proporciones de estudiantes de los diferentes niveles que están a favor del cambio en el plan de estudios.

Como una ilustración más de esta función de Excel, en el ejemplo 11.8 se probó si los montos de 220 facturas se ajustaban a una distribución normal. Después de realizadas las operaciones, se encontró una χ^2 calculada de 13.7 que resultó ser mayor que la χ^2 crítica, así que se rechazó la hipótesis nula de que los montos de las facturas se ajustaban a una distribución normal. Ahora, con la función =DISTR.CHI(13.7,6) se obtiene una probabilidad de 0.033 con el nivel de significación establecido de 5% (0.05); esa probabilidad de 0.033 es menor que 0.05 por lo que se rechaza, al igual que antes, que la distribución de los montos de las facturas se ajuste a una distribución normal.

11.14.2 La función Prueba.Chi

La función Prueba.Chi también es muy útil, es posible que lo sea aún más que las otras 2, considerando que ahorra la laboriosa tarea de los cálculos de:

$$\frac{(f_0 - f_e)^2}{f_e}$$

Las operaciones de esta fórmula son necesarias para determinar el valor calculado de la ji cuadrada y arroja la probabilidad de obtener una ji cuadrada calculada como la que producen los datos, mismos que Excel no revela con esta función.

La sintaxis de esta función es:

Prueba.Chi(rango de frecuencias observadas, rango de frecuencias esperadas)

En el ejemplo 11.6 se probó si eran iguales o no las proporciones de artículos defectuosos provenientes de 2 procesos de producción diferentes. Con un nivel de significación de 1%. De los datos observados se determinaron los esperados y se llegó a las frecuencias que se muestran en la tabla 11.12.

Tabla 11.12 Frecuencias observadas y esperadas para los datos del ejemplo 11.6

	Proceso 1	Proceso 2	Proceso 1	Proceso 2
	f_o	f_o	f_e	f_e
Defectuosos	35	20	30.55	24.45
No defectuosos	465	380	469.45	375.55

Si los datos de las frecuencias se colocan en las celdas B3:E4, la función =PRUEBA.CHI(B3:C4,D3:E4) produce como resultado 0.2127, es decir, la probabilidad de obtener una ji cuadrada calculada (que la función no revela) como la que se desprende de estos datos.

Al observar el valor de probabilidad de 0.2127 puede concluirse que la hipótesis nula se acepta, dado que 21.27% es un valor de probabilidad muy superior a cualquiera de los valores que suelen utilizarse para el nivel de significación (los más comunes suelen ser 1 o 5%).

El ejemplo anterior ilustra que esta función de Excel puede ser útil para ahorrar los laboriosos cálculos de la ji cuadrada. Sin embargo, debe utilizarse con cautela porque un usuario con poca experiencia en pruebas de hipótesis con χ^2 fácilmente podría cometer errores de interpretación si no está consciente de los cálculos y el procedimiento.

11.15 Resumen

Se analizan diversas pruebas con la distribución ji cuadrada como distribución de probabilidad teórica.

En las pruebas para la varianza de una población, el estadístico de prueba χ^2 es:

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} \qquad (11.3)$$

Con $n - 1$ grados de libertad.

En las pruebas de hipótesis restantes: una, 2 o n proporciones; de bondad de ajuste o de independencia y homogeneidad, el cálculo se hace a través de frecuencias observadas y frecuencias esperadas:

$$\chi^2 = \sum\left[\frac{(O_i - E_i)^2}{E_i}\right] \qquad (11.4)$$

Con $n - k - 1$ grados de libertad.

Las pruebas para una o 2 proporciones pueden realizarse indistintamente utilizando como estadístico de prueba la z de la distribución normal o la χ^2, ya que son equivalentes y producen los mismos resultados. Finalmente, se marcó la diferencia entre pruebas paramétricas y no paramétricas para señalar que en este capítulo se hicieron pruebas de ambos tipos.

11.16 Fórmulas del capítulo

11.2 Distribución ji cuadrada (χ^2)

Función de densidad de probabilidad de la distribución χ^2:

$$f(u) = \frac{1}{\left(\frac{n}{2} - 1\right)!\,2^{\frac{n}{2}}} u^{\frac{n}{2}-1} e^{-\frac{u}{2}}, u > 0 \qquad (11.1)$$

En donde:

$$u = \sum_{i=1}^{n}\left[\frac{(X_i - \mu_i)^2}{\sigma_i^2}\right] \qquad (11.2)$$

Con $n - k - 1$ grados de libertad.

11.4 Pruebas de hipótesis para la varianza de una población

Estadístico de prueba ji cuadrada, χ^2, para la varianza de una población:

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} \qquad (11.3)$$

Con $n - 1$ grados de libertad.

11.5 Distribución ji cuadrada a partir de frecuencias observadas y frecuencias esperadas

$$\chi^2 = \sum\left[\frac{(O_i - E_i)^2}{E_i}\right] \qquad (11.4)$$

Con $n - k - 1$ grados de libertad.

11.17 Ejercicios adicionales

11.4 Pruebas de hipótesis para la varianza de una población

1. Una empresa manufacturera tiene un pedido de 15 mil playeras que debe entregar en 3 meses. Para el cliente es muy importante que la cantidad de poliéster sea de 30 g y que la varianza no sea superior a 4 g^2, por lo que cada 15 días hará una revisión en las playeras producidas. En la primer revisión se toma una muestra aleatoria de 300 playeras donde la varianza del contenido de poliéster resulta ser de 3.5. Determine, con un nivel de significación de 0.05, si se está cumpliendo con los requerimientos del pedido.
2. Una fábrica de materiales industriales ha lanzado un nuevo producto cuyo diámetro debe ser de 30 cm con una varianza de 0.07 para reemplazar la línea anterior sin ninguna complicación. Se está utilizando la misma maquinaria que se emplea para el resto de la producción, por lo que el jefe de operación está muy preocupado en cuanto al cumplimiento de la medida de la varianza señalada. Para ello, toma una muestra aleatoria de 50 de los nuevos productos y obtiene que su varianza es de 0.065. Indique, con un nivel de significación de 0.025, si se está cumpliendo con esta medida.
3. El señor Yáñez tiene como objetivo recuperar su inversión en 240 días con una varianza de 30. Hace un par de años que no evalúa estas condiciones, así que toma una muestra aleatoria de 8 inversiones y observa que la varianza del tiempo de recuperación de ellas es de 35. Determine con un nivel de significación de 0.5, si la varianza ha cambiado.
4. Una mueblería encarga a un nuevo proveedor piezas de madera cuyo grosor debe ser de 20 cm con una varianza de 2.5 mm^2. Al recibir las piezas el encargado de almacén de la mueblería toma una muestra aleatoria de 30 piezas y encuentra que tienen un grosor de 20 cm con una varianza de 2.2. Determine con un nivel de significación de 0.01 si el nuevo material cumple con las especificaciones.
5. Una empresa de tecnología tiene establecida una inversión promedio de $1 500 000 con una varianza de $100 000 por cada investigación. El área de diseño del producto ha propuesto una serie de proyectos de investigación para el próximo periodo; una muestra aleatoria de estos 3 refleja que la inversión necesaria por cada uno sería aproximadamente de $1 700 000 con una varianza de $150 000. Con un nivel de significación de 0.5, determine si la varianza ha cambiado y decida si es viable aceptar o no dichos proyectos.

11.6 Prueba para una proporción con z y con χ^2

6. Una empresa dedicada a realizar investigaciones de mercado afirma que cuando mucho 10% de los grandes empresarios mexicanos invierten parte de su dinero en el banco. Para probar esto, se toma una muestra de 45 empresarios y se observa que 15 de ellos tienen inversiones en algún banco. ¿Puede rechazar la afirmación de la empresa, con un nivel de significación de uno por ciento?
7. Una empresa manufacturera de ropa asegura que 15% de su producción presenta algún defecto. Para demostrarlo, se toma una muestra de 270 artículos y se encuentra que 40 presentan algún defecto. ¿Puede rechazar la afirmación de la empresa, con un nivel de significación de 1.5 por ciento?
8. Una empresa de investigación afirma que 35% de los ciudadanos prefieren al partido político A, mientras que el resto se divide entre los demás partidos. Para confirmar estos datos se toma una muestra de 125 ciudadanos, de los cuales 40 prefieren al partido A. ¿Puede rechazar la afirmación de la empresa de investigación, con un nivel de significación de uno por ciento?
9. El Instituto de Oftalmología afirma, con base en una encuesta realizada, que 30% de la población menor a 10 años de edad presenta algún problema de la vista. Para probarlo se toma una muestra de 70 niños, de los cuales 25 presentaron algún defecto de este tipo. ¿Puede rechazar la afirmación del instituto, con un nivel de significación de uno por ciento?
10. Una de las más importantes marcas de tequila del país cuenta con datos que afirman que 50% de los consumidores de esta bebida prefieren su marca por encima de las demás. Para reafirmar esta información se toma una muestra de 110 consumidores de esta bebida y 80 de ellos prefieren su tequila. ¿Puede rechazar esta afirmación, con un nivel de significación de uno por ciento?

11.7 Prueba para la diferencia entre 2 proporciones con z y con χ^2

11. Un estudio de mercadotecnia refleja mediante una muestra aleatoria que 35 de 50 niños de una ciudad del norte y 32 de 50 de una del sur, prefieren videojuegos de una marca específica. Pruebe la hipótesis de que estas proporciones son iguales con un nivel de significación de 2.5 por ciento.
12. Un empresario tomó una muestra aleatoria de 100 empleados del área administrativa y otra de 500 de la operativa. Encontró que 8 y 25, respectivamente, no están cumpliendo con su trabajo. Determine con un nivel de significación de 1% si la hipótesis de que las proporciones de trabajadores con esta característica es la misma en las 2 áreas.
13. Una empresa de telecomunicaciones toma una muestra aleatoria de 70 llamadas locales, de las cuales 20 tienen una duración superior a 15 minutos. Otra muestra aleatoria de 50 llamadas internacionales refleja que sólo 15 rebasan este tiempo. Tomando en cuenta un nivel de significación de 5%, pruebe la hipótesis de que la proporción de llamadas que duran más de 15 minutos entre locales e internacionales es la misma.
14. Un laboratorio toma una muestra aleatoria de 150 pacientes hombres y 200 mujeres a quienes se les aplicó un tratamiento. Al cabo de 10 días, 100 hombres y 180 mujeres muestran mejoría gracias al tratamiento. Con un nivel de significación de 1% determine si la proporción es la misma para hombres y mujeres que tuvieron una buena reacción.
15. Un banco compara 2 sucursales esperando que la proporción de transacciones que generan comisión alta sea la misma. De una muestra aleatoria de 600 transacciones de la

sucursal 1 270 generan una comisión alta; de otra muestra aleatoria de 720 transacciones de la sucursal 2 340 cumplen con esta característica. Pruebe con un nivel de significación de 0.5% la hipótesis de que esta proporción es la misma en las 2 sucursales.

11.9 Prueba para la diferencia entre *n* proporciones

16. Para evaluar si el nivel educativo está relacionado con las preferencias políticas, se tomó una muestra de ciudadanos en edad de votar a fin de preguntarles sobre sus preferencias en términos de partidos políticos. Se obtuvieron los siguientes resultados:

Partido	Nivel educativo			
	Primaria	Secundaria	Media superior	Superior
PRI	50	80	60	40
PAN	20	15	20	10
PRD	40	70	60	50
Otro	10	10	20	12
Ninguno	30	35	20	40

Pruebe si estas 2 variables son independientes a un nivel de significación de 0.05.

17. Las matemáticas son una asignatura obligatoria en las 3 licenciaturas que se imparten en la Facultad de Administración. Se toma una muestra aleatoria para evaluar si la calificación de los alumnos en el curso básico es independiente de la licenciatura que han elegido. Los resultados son los siguientes:

Calificación	Licenciatura elegida		
	Administración	Contaduría	Informática
6 o menos	50	60	4
7 u 8	286	140	41
9 o 10	92	70	9

Pruebe si estas 2 variables son independientes, con un nivel de significación de 0.01.

18. Se realizó una encuesta para determinar si la edad está relacionada con la opinión de los ciudadanos sobre el tipo de seguridad social para los trabajadores, y se obtuvieron los siguientes resultados:

	Grupo de edad		
Tipo de seguridad social	**18-30**	**31-60**	**Más de 60**
Garantizada por el Estado	92	137	72
Con fondos de pensiones	61	114	56
Sin opinión	19	73	3

Pruebe si estas 2 variables son independientes con un nivel de significación de 0.05.

19. ¿La edad de muerte a consecuencia de diabetes depende del sexo? Utilice los datos siguientes para probar esta hipótesis con un nivel de significación de 0.01.

Grupo de edad	Sexo	
	Hombre	Mujer
0 a menos de 20	3	2
20 a menos de 40	5	3
40 a menos de 50	11	7
50 a menos de 60	21	13
60 o más	72	88

20. Con los datos siguientes, pruebe si el estado civil es independiente de la edad con un nivel de significación de 0.05.

Estado civil	Edad				
	18 a menos de 25	25 a menos de 35	35 a menos de 45	45 a menos de 55	55 o más
Casado	24	66	92	95	81
Soltero	88	95	88	94	121
Divorciado	15	74	113	115	63
Viudo	1	18	30	90	184

11.10 Pruebas de bondad de ajuste a distribuciones teóricas. Pruebas de bondad de ajuste a una distribución normal

21. El administrador de una página web desea probar si la distribución del número de visitas diarias se ajusta a una distribución normal con un nivel de significación de 1%. Con una muestra aleatoria de 40 días se obtienen los siguientes datos:

Núm. de visitas diarias, X_i	F_o
0 a menos de 250	1
250 a menos de 500	10
500 a menos de 750	6
750 a menos de 1 000	5
1 000 a menos de 1 250	7
1 250 a menos de 1 500	4
1 500 a menos de 1 750	5
1 750 a menos de 2 000	2
Totales	40

22. El jefe de operaciones de una fábrica obtiene una muestra aleatoria de 35 días en los que se produjeron artículos y quiere saber si el nivel de producción se ajusta a una distribución normal con un nivel de significación de 0.05.

Los datos de la muestra se presentan a continuación:

Núm. de artículos producidos en un día, X_i	f_o
0 a menos de 3 500	2
3 500 a menos de 7 000	9
7 000 a menos de 10 500	6
10 500 a menos de 14 000	4
14 000 a menos de 17 500	6

(*continúa*)

(*continuación*)

Núm. de artículos producidos en un día, X_i	f_o
17 500 a menos de 21 000	8
Totales	35

23. El gerente de un banco desea saber si el monto de las transacciones que se realizan se ajusta a una distribución normal con un nivel de significación de 0.025. Toma una muestra aleatoria de 5 000 transacciones y obtiene los siguientes resultados:

Monto de la transacción, X_i	f_o
0 a menos de 1 500	1 500
1 500 a menos de 3 000	1 300
3 000 a menos de 4 500	470
4 500 a menos de 6 000	500
6 000 a menos de 7 500	300
7 500 a menos de 9 000	200
9 000 a menos de 10 500	220
10 500 a menos de 12 000	100
12 000 a menos de 13 500	80
13 500 a menos de 15 000	60
15 000 en adelante	270
Totales	5 000

11.9.2 Pruebas de bondad de ajuste a una distribución Poisson

24. Se desea saber cuántas personas entran a un edificio ejecutivo en un intervalo de 5 min, de tal manera que se toma una muestra de 70 intervalos y se obtienen los resultados que se muestran a continuación. Con un nivel de significación de 0.01, pruebe si la distribución del número de personas que entran al edificio sigue una distribución Poisson.

Núm. de personas que entran, X	Frecuencias observadas, f_o	$X \cdot f_o$
0	5	0
1	6	6
2	8	16
3	12	36
4	9	36
5	10	50
6 o más	20	120
	70	264

25. Pruebe, con un nivel de significación de 0.025, si el número de demoras que presenta diariamente el transporte colectivo sigue una distribución Poisson. Para ello, se toma una muestra de 30 días y se obtienen los siguientes datos:

Núm. de atrasos en el sistema, X	Frecuencias observadas, f_o	$X \cdot f_o$
0	5	0
1	15	15
2	3	6
3	2	6
4	3	12
5 o más	2	10
	30	49

26. Pruebe si el número de artículos defectuosos, que genera a diario una máquina de tejido, sigue una distribución de Poisson. Se toma una muestra de 150 días y se obtienen los siguientes resultados:

Núm. de artículos defectuosos, X	Frecuencias observadas, f_o	$X \cdot f_o$
0	0	0
1	15	15
2	10	20
3	12	36
4	9	36
5	10	50
6	19	114
7	18	126
8	20	160
9	17	153
10 o más	20	200
	150	910

Realice la prueba con un nivel de significación de 0.05.

27. Se desea saber si el número diario de llamadas que entran a una estación de bomberos sigue una distribución de Poisson. Se toma una muestra aleatoria de 65 días que arrojó los siguientes resultados:

Núm. de llamadas entrantes, X	Frecuencias observadas, f_o	$X \cdot f_o$
0	0	0
1	7	7
2	6	12
3	5	15
4	3	12
5	5	25
6	5	30
7	4	28
8	8	64
9	7	63
10 o más	15	150
	65	406

Realice la prueba con un nivel de significación de 0.04.

11.9.3 Pruebas de bondad de ajuste a una distribución binomial

28. El gerente realiza una evaluación donde relaciona el número de ventas diarias por empleado suponiendo que ésta se asemeja a una distribución binomial. Con una muestra de 40 días obtiene los siguientes resultados:

Promedio del núm. de ventas por empleado en un día	Núm. de días
0	4
1	9
2	9
3	7
4	6
5	3
6 o más	2

Determine con un nivel de significación de 0.025 si realmente esta relación se ajusta a una distribución binomial, sabiendo que la empresa tiene 15 ventas diarias como mínimo.

29. Durante un largo periodo, el gerente de un despacho de contabilidad realiza diariamente una evaluación a 5 becarios sobre conocimientos en impuestos; con una muestra aleatoria de 25 días encuentra los siguientes resultados:

Núm. de becarios que aprobaron el examen diariamente	Núm. de días
0	4
1	4
2	7
3 o más	10

Determine con un nivel de significación de 0.005 si los datos se adecuan a una distribución binomial.

30. El jefe de operaciones de una fábrica de aparatos electrónicos toma una muestra aleatoria de 40 días en los que evalúa los productos defectuosos y obtiene los siguientes datos:

Núm. de productos defectuosos	Núm. de días
0	3
1	3
2	12
3	12
4	4
5	4
6	2
7 o más	0

Determine con un nivel de significación de 0.05 si los datos se ajustan a una distribución binomial, considerando que la empresa cuenta con 20 máquinas.

31. El director de un proyecto de construcción toma una muestra aleatoria de 35 días para evaluar el avance diario; toma en cuenta el número de etapas terminadas en un día. Obtiene los siguientes resultados:

Núm. de etapas terminadas	Núm. de días
0	6
1	9
2	9
3	8
4	3
5 o más	0

Determine con un nivel de significación de 0.025 si los datos presentados se ajustan a una distribución binomial. Tome en cuenta que la construcción está dividida en 25 etapas.

11.11 Pruebas de bondad de ajuste a distribuciones empíricas

32. Una cadena de tiendas de autoservicio desea probar si el detergente de su propia marca se vende igual que los de otras marcas. Para hacerlo, toma una muestra de 500 ventas de detergente en sus diversas tiendas. Encuentra que de la marca A se vendieron 120 bolsas; de la B, 150; de la C, 125, y de la marca propia, 105 unidades. ¿Puede afirmar, con un nivel de confianza de 0.01, que la venta de estos 4 detergentes es pareja?

33. En una escuela de estudios superiores se desea investigar si las calificaciones de estudiantes a profesores es igual en todas las áreas. Se selecciona una muestra aleatoria de evaluaciones de las distintas áreas. Los resultados se muestran a continuación:

Área académica	Calificaciones			
	Excelente	Bueno	Regular	Malo
Administración	21	28	15	8
Contaduría	18	25	12	7
Economía	15	21	13	6

Pruebe si la distribución de las calificaciones entre las tres areas académicas es igual o no, con un nivel de significación de 0.05.

34. Un investigador intenta determinar si la eficiencia de los subordinados de los administradores varía de acuerdo con los diferentes estilos para dirigir. Ha clasificado los estilos de administración en 3 categorías: tradicional, democrático e innovador. Toma una muestra de trabajadores que laboran bajo la dirección de administradores con los 3 estilos y encuentra los resultados siguientes:

Desempeño de los subordinados	Estilo de dirección		
	Tradicional	Democrá-tico	Innovador
Superior al promedio	40	60	80
Promedio	80	180	120
Inferior al promedio	20	10	20

Pruebe la suposición del investigador de que la distribución de los desempeños de los subordinados es la misma para

todos los estilos de administración con un nivel de significación de 0.01.

11.12 Pruebas sobre la independencia entre 2 variables

35. Una empresa de telefonía celular desea saber si el modelo de teléfono que prefieren sus clientes depende de la edad de éstos. Para saberlo, toma una muestra aleatoria de 500 compradores, se identifican los modelos y las edades de los clientes. Pruebe si las 2 variables son independientes con un nivel de significación de 0.05.

Edades	Modelo de celular					
	1	2	3	4	5	Total
15-25	10	28	16	20	55	129
26-36	25	13	27	20	19	104
37-47	7	19	40	10	30	126
48 +	28	15	22	30	46	141
Total	90	75	105	80	150	500

36. El director de una aseguradora desea saber si el tipo de seguros depende de los ingresos del cliente. Para investigarlo, toma una muestra aleatoria de 250 personas e identifica los tipos de seguros que prefieren y cuáles son sus ingresos. Pruebe la independencia de estas 2 variables con un nivel de significación de 0.10.

Ingresos	Tipos de seguro			
	A	B	C	Total
0-$5 000	15	10	11	36
$5 001-$15 000	20	29	19	68
$15 001-$30 000	15	38	10	63
$30 001-adelante	20	23	40	83
Total	70	100	80	250

37. El coordinador de una universidad desea saber si las calificaciones de los alumnos de la licenciatura a su cargo dependen del género de los estudiantes. Para saber esto, toma una muestra de 200 alumnos e identifica las calificaciones de hombres y mujeres. Con un nivel de significación de 0.025, ¿las calificaciones y el género son variables independientes?

Sexo	Calificaciones		
	Aprobatoria	Reprobatoria	Total
Femenino	53	43	96
Masculino	42	62	104
Total	95	105	200

38. Una empresa cervecera desea saber si el tipo de cerveza preferida por sus clientes (ligera, clara, oscura) depende del género del consumidor. Para saber esto, toma una muestra de 150 personas e identifica los tipos de cerveza y si sus consumidores son hombres o mujeres. Pruebe si el género es independiente al tipo de cerveza seleccionada, con un nivel de significación de 0.05.

Género	Tipos de cerveza			
	Ligera	Clara	Oscura	Total
Hombre	22	25	19	66
Mujer	28	35	21	84
Total	50	60	40	150

39. Una institución gubernamental desea saber si el desempeño académico está relacionado con el estado nutricional de los niños. Para saber esto, toma una muestra de 300 niños e identifica cuántas comidas completas ingieren al día y qué nivel de desempeño académico mantienen. Pruebe, con un nivel de significación de 0.005, si el nivel nutricional y el desempeño académico son variables independientes.

Desempeño	Comidas completas			
	1	2	3	Total
Bueno	20	60	80	160
Malo	50	60	30	140
Total	70	120	110	300

capítulo 12

Análisis de varianza

Sumario

12.1 Introducción
12.2 Suposiciones en que se basan las técnicas de análisis de varianza
12.3 El diseño completamente aleatorizado de 1 factor
12.4 Procedimiento para el ANOVA con el diseño completamente aleatorizado de 1 factor
12.5 Excel y ANOVA de un factor
12.6 Comparaciones múltiples entre pares de medias de tratamiento
12.7 Análisis de varianza de 2 factores
12.8 Excel y ANOVA de 2 factores
12.9 Análisis de varianza de 2 factores con interacción
12.10 Excel y ANOVA de 2 factores con interacción
12.11 Resumen
12.12 Fórmulas del capítulo
12.13 Ejercicios adicionales

12.1 Introducción

Se conoce como **análisis de varianza**, o ANOVA por sus siglas en inglés (*ANalysis Of VAriance*), a un conjunto de técnicas que se utilizan para probar hipótesis sobre la igualdad de más de 2 medias, que es un tema que extiende los procedimientos de pruebas de hipótesis para 1 o para 2 medias que se analizaron antes, particularmente en el capítulo 10, que trata sobre pruebas de hipótesis para 2 poblaciones.

Análisis de varianza (ANOVA). Es un conjunto de técnicas que se utilizan para probar hipótesis sobre la igualdad de más de 2 medias.

En ese capítulo se revisará el procedimiento para realizar una prueba para la diferencia entre 2 varianzas utilizando como estadístico de prueba a la F de Fisher, que es también el estadístico de prueba en ANOVA.

Se pueden aplicar estas técnicas de pruebas de hipótesis para más de 2 medias en muchos campos. Por ejemplo:

- Un gerente de compras (o abastecimientos) puede interesarse en comparar la durabilidad media de las llantas de 3 o más proveedores de su empresa.
- A un gerente de producción (o de operaciones) le interesa saber si diferentes procesos productivos tienen el mismo rendimiento, o existen diferencias entre ellos.
- Un gerente de mercadotecnia quiere saber si la productividad por visita de sus vendedores es la misma o no.
- Un analista financiero desea saber si el margen de operación de diversas empresas es el mismo, o difiere.

El análisis de varianza se utiliza, como se menciona antes, para probar la igualdad de más de 2 medias, pero las pruebas se llevan a cabo midiendo las diferencias entre las distintas varianzas de las poblaciones; es por ello que a esta técnica se le conoce con el nombre de análisis de *varianza*, aunque, más explícitamente, se utilizan las sumas de cuadrados (SC) de diferencias entre los datos y sus medias, $\sum(X - \bar{X})^2$, que son el numerador de donde se obtiene la varianza de una muestra, después de que se divide esta SC entre $n - 1$:

$$S^2 = \frac{\sum(X - \bar{X})^2}{n - 1}$$

Como se verá más adelante, se utilizan diversas sumas de cuadrados divididas entre distintos grados de libertad (que son, en cierta forma, varianzas) para calcular el estadístico con el que se realizan las pruebas, que es la F de Fisher que ya se usó antes.

En la sección siguiente se presenta una introducción al tema, junto con algunos conceptos importantes y, más adelante, se tiene otra sección en la que se analizan los supuestos que se deben cumplir para que las conclusiones del análisis sean válidas.

En las secciones subsiguientes se presentan los principales procedimientos del análisis de varianza:

1. Análisis de varianza de un factor, junto con una extensión de este análisis que se ocupa de hacer comparaciones múltiples entre pares de medias de tratamiento.
2. Análisis de varianza de 2 factores.
3. Análisis de varianza de 2 factores con interacción.

Como se pretende que el lector comprenda cabalmente la metodología de análisis, se resuelven ejercicios paso a paso pero, también, se incluyen 3 secciones en donde se explican 3 herramientas de Excel que facilitan enormemente los laboriosos cálculos que se requieren para esos tipos de análisis de varianza.

La forma más sencilla de visualizar los diferentes tipos de análisis de varianza es observando sus nombres:

1. Análisis de varianza de un factor.
2. Análisis de varianza de 2 factores.
3. Análisis de varianza de 2 factores con interacción.

En los 3 ejemplos siguientes se ilustran estos 3 casos, comenzando con el ejemplo 12.1, en el que se presenta una situación típica de ANOVA de un factor.

EJEMPLO 12.1

Una empresa ensambla cuadros para un solo modelo de bicicleta. La planta trabaja 3 turnos: el matutino, el vespertino y el nocturno; cada mes los trabajadores se rotan, por lo que al cabo de un trimestre todos ellos laboraron en los tres turnos; el director de producción quisiera saber si existe diferencia entre la producción promedio de los 3 turnos, ya que la fábrica trabaja a plena capacidad desde hace más de un año y no hay variaciones significativas en el número de empleados, en la maquinaria ni en los procesos productivos, el director considera que con un análisis de varianza puede contestar esta pregunta. Para ello, obtiene una muestra aleatoria de la producción de 6 días de cada turno, los datos se muestran a continuación.

Turno		
Matutino	**Vespertino**	**Nocturno**
129	138	118
141	142	120
128	140	132
145	149	118
135	129	136
144	148	138

Se plantean las siguientes hipótesis:

$$H_0: \mu_1 = \mu_2 = \mu_3$$

H_1: Por lo menos una de las medias poblacionales no es igual a las otras.

En donde:

μ_1 = Producción media en el turno matutino.
μ_2 = Producción media en el turno vespertino.
μ_3 = Producción media en el turno nocturno.

El factor que se analiza aquí es la producción por turno y los datos de cada uno de ellos dispuestos en sendas columnas. En este análisis de varianza de un factor se utilizan abundantemente 2 conceptos: **tratamiento** y **unidad experimental**: el primero se refiere a cualquier condición que se controla en el experimento como, por ejemplo, el medicamento que se administra a diversas personas o animales, el tipo de fertilizante que se utiliza en distintos sembradíos o el nivel de presión que se aplica en el mecanismo que llena las botellas con algún líquido; en el ejemplo, el tratamiento es el turno trabajado; por su parte, se denomina *unidad experimental* a cada uno de los sujetos (persona, animal, corrida, sembrado o botella, por ejemplo) a los que se les aplica determinado tratamiento; en el ejemplo, la unidad experimental es cada turno en cada día.

Tratamiento. Es cualquier condición que se controla en el experimento.
Unidad experimental. Cada uno de los sujetos (persona, animal, corrida, sembrado o botella, por ejemplo) a los que se les aplica determinado tratamiento.

Por otra parte, se tiene la **variable de respuesta**, que es la medición que se realiza sobre las unidades experimentales. En el caso del análisis de distintos fertilizantes, la variable de respuesta es la cantidad de producción, que puede ser el promedio de toneladas de maíz por hectárea, por ejemplo. En el caso de los diferentes niveles de presión para el llenado de botellas, la variable de respuesta puede ser el promedio de contenido de líquido de las botellas llenadas; en el ejemplo, la variable de respuesta es la producción por día y por turno, es decir, el número de cuadros de bicicleta ensamblados en cada turno de cada día.

Variable de respuesta. Es la medición que se realiza sobre las unidades experimentales.

Así, las 2 variables que se manejan en este ejemplo son: el turno, que es una variable cualitativa ordinal y la producción de cuadros, que es una variable numérica. Para este ejemplo, se considera que la variable dependiente o de respuesta es la producción por turno (unidades producidas), ya que se trata de averiguar si el volumen de producción depende del turno en que se labora; por otra parte, la variable independiente o de tratamiento es el turno en que dicha producción se lleva a cabo.

Por su parte, el ANOVA de 2 factores se aplica cuando se tienen datos clasificados de acuerdo con *2* variables, como en el ejemplo siguiente.

Ejemplo 12.2

Una pequeña empresa que elabora botanas (papas fritas, cacahuates, etc.) tiene 4 rutas para el reparto de sus productos dentro de la ciudad. Para cubrir estas rutas tiene 4 conductores básicos y 1 complementario, el conductor complementario sirve para cubrir ausencias por vacaciones o enfermedad de los otros 4 y se le ocupa para otros menesteres si no se le requiere como conductor. El puesto de conductor complementario se rota entre todos los choferes sobre una base mensual. Es importante, tanto desde el punto de vista de eficiencia, como de justicia, que todas las rutas tengan la misma duración, también es importante que todos los conductores sean igualmente eficientes en todas las rutas, para que los conductores puedan sustituirse el uno al otro sin pérdida de tiempo; por ello, la gerencia de distribución implementó un programa de capacitación para que los conductores se familiaricen con todas las rutas. Al cabo de esta capacitación se hicieron recorridos de prueba, con el objeto de investigar, por una parte, si todas las rutas se recorrían en el mismo tiempo y, por otra, si los conductores eran igualmente eficientes en todas. La gerencia de distribución desea saber si el tiempo medio de las cuatro rutas es igual y si la eficiencia media de todos los conductores en las rutas es igual. La información de los recorridos y los conductores se da a continuación:

Conductor/ruta	Ruta A	Ruta B	Ruta C	Ruta D
Antúnez	224	227	237	248
Becerra	242	235	262	250
Cervantes	225	240	235	261
Domínguez	232	253	259	255
Escamilla	232	245	257	261

El tiempo se mide en minutos. **NOTA**

En este ejemplo se tienen también datos ordenados en columnas pero, a diferencia del ejemplo 12.1, los datos se clasifican de acuerdo con 2 variables: el conductor y la ruta que son, entonces, los 2 factores. Aquí, las hipótesis que se prueban son:

Para las rutas:

$H_0: \mu_A = \mu_B = \mu_C = \mu_D$
H_1: Por lo menos una de las medias poblacionales de los tratamientos no es igual a las otras.

En donde las diferentes medias son las medias de las 4 rutas.

Para los conductores:

$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$
H_1: Por lo menos una de las medias poblacionales de los bloques (conductores) no es igual a las otras.

En donde las diferentes medias son las medias de los 5 conductores.

En este ejemplo, el tratamiento son las diferentes rutas y a la variable del conductor, cuyos datos están en los renglones, se le conoce como *grupos*; a su vez, tratamientos y grupos son los 2 factores a los que se refiere el nombre de esta técnica de análisis de varianza de 2 factores.

Un detalle que vale la pena tener presente respecto a este ANOVA de 2 factores es que se tiene un solo dato para cada conductor y para cada ruta, ya que el análisis de varianza de este par de factores, con interacción que se revisa más adelante, se refiere también a conjuntos de datos clasificados de acuerdo con 2 variables, pero en este caso, con cuando menos 2 observaciones para cada par de factores. Se ilustra esto en el ejemplo siguiente:

Ejemplo 12.3

Una empresa que fabrica y vende purificadores de agua para el comercio y la industria tiene 3 zonas de ventas: centro, norte y sur. Como la labor de venta tiene aspectos tanto técnicos como financieros, la empresa tiene 3 vendedores: un químico, un licenciado en administración y un ingeniero mecánico. La gerencia de la empresa muestra interés en saber si las 3 zonas tienen un potencial equivalente si los 3 vendedores tienen igual capacidad y si es indistinto el trabajo de los vendedores en cualquier zona o existen diferencias. Esta ampliación de nuestra perspectiva requiere un poco más de información, es necesario contar con al menos 2 mediciones para cada combinación vendedor-zona; es decir, se tiene que replicar la observación, si se tienen 2 observaciones para cada combinación vendedor-zona, se dice que se tienen 2 réplicas; si se tiene 3 observaciones, 3 réplicas y así sucesivamente. Las variables independientes son los factores: zonas y vendedores, la variable dependiente es el volumen de ventas. La información pertinente se da en la siguiente tabla. Se verá que hay 2 datos para cada combinación zona-vendedor;

para esta técnica es indispensable tener por lo menos 2 observaciones para cada combinación de factores y el número de observaciones para cada una de estas combinaciones debe ser igual, son 2 observaciones para cada una en nuestro ejemplo, pero pueden ser 3 o más, siempre que sea el mismo número para todas.

	Ventas mensuales en miles de pesos		
Profesión del vendedor	**Zona centro**	**Zona sur**	**Zona norte**
Químico	506	528	513
	512	534	495
Licenciado en administración	529	496	508
	525	498	500
Ingeniero mecánico	500	512	528
	518	504	520

Se identifican 3 variables: las zonas, que se manejaron como variable de tratamiento y que ahora van a ser uno de los factores, el factor A; la profesión de los vendedores, a la que se le llamó *variable de bloque* y que ahora será el factor B; se tiene una tercera variable, que fue dada por la interacción entre los factores A y B y que se mide, precisamente, a través de los pares de datos que se tienen para cada combinación de zona (factor A) y de vendedor (factor B).

Para probar las hipótesis de ANOVA se utilizará un método de estadístico de prueba de una sola cola, de una manera similar a la que se trabajó en el capítulo 10. Para ello, se utiliza la misma distribución F de Fisher ya vista y se calcula un valor de F, con base en los datos observados: la F empírica, la cual se compara con el valor teórico o crítico que se obtiene de la tabla para un determinado nivel de confianza. Si el valor empírico no rebasa el valor crítico, no se rechaza la hipótesis nula; en caso contrario, se le rechaza y se acepta la hipótesis alternativa.

La manera de probar la hipótesis de igualdad de las medias consiste en obtener diferentes sumas de cuadrados (SC) y, con éstas, obtener promedios de cuadrados (PC), luego se obtiene la razón, o cociente, de los 2 promedios de cuadrados. Si esa razón se acerca a la unidad (o no se aleja demasiado de la unidad) se concluirá que todas las medias son iguales; en caso contrario, se rechazará esta suposición. En qué medida se puede alejar de la unidad la razón de los promedios de cuadrados (el valor empírico de F) sin que se rechace H_0 depende, por supuesto, del valor crítico de F, de acuerdo con el valor que se identifica en la tabla de esa distribución F.

Esas sumas y promedios de cuadrados difieren según el modelo de ANOVA aplicado; en las secciones 12.3 y siguientes se ilustran las diferentes técnicas.

12.2 Suposiciones en que se basan las técnicas de análisis de varianza

Tal como sucede con otras técnicas que ya se revisaron antes, el análisis de varianza se basa en un conjunto de supuestos que deben cumplirse para que las conclusiones sean válidas y, también igual que antes, estos supuestos rara vez se cumplen a cabalidad, por lo que debe siempre evaluarse su cumplimiento tomando las conclusiones obtenidas como resultados aproximados y no absolutos. Existe un supuesto en el que se basan todos los modelos del análisis de varianza y que es: *la variable de respuesta tiene distribución normal en cada una de las poblaciones.* Además, existen otras suposiciones que varían según el diseño utilizado. En cada sección se especifican las suposiciones aplicables al modelo que se estudia.

Los procedimientos de ANOVA pueden servir para analizar datos procedentes tanto de observaciones simples o de encuestas como provenientes de experimentos específicamente diseñados, y pueden servir también para analizar situaciones con diferentes números de variables (diseños de 1 o de 2 factores) y con diferentes enfoques (diseños con y sin interacción). Estas condiciones dan lugar a diferentes diseños; los que se estudian en el presente capítulo son:

1. Diseño completamente aleatorizado de 1 factor.
2. Diseño completamente aleatorizado de 2 factores.
3. Diseño completamente aleatorizado de 2 factores con interacción.
4. Diseño de cuadrados latinos.

En las secciones siguientes se revisa la metodología para realizar pruebas de hipótesis para más de 2 medias según cada uno de estos tipos de diseños de análisis de varianza y sólo se intercala una sección adicional inmediatamente después del diseño completamente aleatorizado de 1 factor, para revisar la forma en la que se llevan a cabo pruebas para el conjunto de todos los pares de medias que se pueden formar en ese diseño de 1 factor.

12.3 El diseño completamente aleatorizado de un factor

Las suposiciones en las que se basa este diseño son las siguientes:

1. La variable de respuesta tiene distribución normal en cada una de las poblaciones.
2. La varianza de la variable de respuesta, σ^2, es igual en todas las poblaciones.
3. Las observaciones son independientes entre sí.

Si el tamaño de muestra es igual para cada grupo, las pruebas de análisis de varianza son razonablemente robustas a la violación del supuesto número 2.

Así, en este diseño completamente aleatorizado de un factor se tienen muestras independientes y el análisis se realiza respecto a una sola variable (de aquí el nombre de *un factor*).

Para la explicación de esta técnica se utiliza el mismo ejemplo 12.1 de la fábrica de cuadros para bicicletas que se mencionó antes y se utiliza un nivel de significación de 0.05.

Ejemplo 12.4

Una empresa ensambla cuadros para un solo modelo de bicicleta. La planta trabaja 3 turnos: el matutino, el vespertino y el nocturno. Los trabajadores se rotan cada uno de los turnos, por lo que, al cabo de un trimestre, todos ellos laboraron en los 3 turnos. El director de producción quisiera saber si existe diferencia entre la producción promedio de los 3 turnos, ya que la fábrica trabaja a plena capacidad desde hace más de un año y no hay variaciones significativas en el número de empleados, en la maquinaria, ni en los procesos productivos; el director considera que con un análisis de varianza puede contestar esta pregunta; para ello, obtiene una muestra aleatoria de la producción de 6 días de cada turno. Los datos se muestran a continuación.

Turno		
Matutino	**Vespertino**	**Nocturno**
129	138	118
141	142	120
128	140	132
145	149	118
135	129	136
144	148	138

Solución:

1. En primer lugar, las hipótesis:

H_0: $\mu_1 = \mu_2 = \mu_3$

H_1: Por lo menos una de las medias poblacionales no es igual a las otras.

2. Ahora se obtienen las medias; tanto la media de cada tratamiento (turno) como la media general de todos los datos:

Media del turno matutino:

$$\bar{X}_1 = \frac{(129 + 141 + 128 + 145 + 135 + 144)}{6} = 137$$

Media del turno vespertino:

$$\bar{X}_2 = \frac{(138 + 142 + 140 + 149 + 129 + 148)}{6} = 141$$

Media del turno nocturno:

$$\bar{X}_3 = \frac{(118 + 120 + 132 + 118 + 136 + 138)}{6} = 127$$

Media global:

$$\bar{X}_g = \frac{(129 + 141 + \cdots + 138)}{18} = 135$$

Nótese que esta media global se puede calcular también como el promedio de las 3 medias de tratamiento:

$$\bar{X}_g = \frac{(137 + 141 + 127)}{3} = 135$$

3. Se obtiene la suma de los cuadrados totales, a la que llamamos SCT, que es la suma de los cuadrados de la diferencia entre cada dato y la media global.

$$SCT = (129 - 135)^2 + \cdots + (144 - 135)^2 + (138 - 135)^2 + \cdots + (148 - 135)^2 + (118 - 138)^2 + \cdots + (138 - 135)^2 = 1\,608$$

4. Se obtiene la suma de cuadrados de las variaciones entre los tratamientos. La identificamos como SCTr; para ello se calcula la media de cada tratamiento menos la media global; luego, se eleva esa diferencia al cuadrado y, finalmente, se multiplican estos cuadrados por el número de elementos de cada tratamiento:

$$SCTr = 6(137 - 135)^2 + 6(141 - 135)^2 + 6(127 - 135)^2 = 624$$

5. Se obtiene la suma de cuadrados de la variación aleatoria, también llamada *suma de cuadrados del error* y es la variación dentro del tratamiento, la identificamos como SCE. Se consigue, dentro de cada tratamiento, con cada dato menos la media del tratamiento y elevando la diferencia al cuadrado.

$$SCE = (129 - 137)^2 + \cdots + (144 - 137)^2 + (138 - 141)^2 + \cdots + (148 - 141)^2 + (118 - 127)^2 + \cdots + (138 - 127)^2 = 984$$

Aquí es importante resaltar el modelo en el que se basa el diseño completamente aleatorizado de un factor que se ilustra con el ejemplo:

$$SCT = SCTr + SCE \qquad \textbf{(12.1)}$$

Con las sumas de cuadrados encontradas hasta aquí se puede verificar que esta igualdad se cumple:

$$1\,608 = 624 + 984$$

6. Se obtiene el promedio de los cuadrados entre tratamientos, lo identificamos como PCTr. Para ello se divide la suma de cuadrados entre tratamientos (SCTr), entre el número de tratamientos menos 1, en este caso 3 – 1, que es el número de grados de libertad en este caso, es decir, 2 grados de libertad. Éste es el primer promedio de cuadrados (varianza), al que se aludió en la sección introductoria:

$$PCTr = \frac{624}{2} = 312$$

7. Se obtiene la media de los cuadrados del error, la identificamos como *PCE*. Para ello se divide la suma de cuadrados del error (*SCE*) entre el número de datos menos el número de tratamientos. En este caso 18 – 3 = 15, que es también el número de grados de libertad en este caso, es decir, 15 grados de libertad. Éste es el segundo promedio de cuadrados al que se aludió en la sección introductoria:

$$PCE = \frac{984}{15} = 65.6$$

8. Se obtiene el valor empírico de *F*, que es el cociente de estos 2 promedios de cuadrados:

$$F_e = \frac{PCTr}{PCE} = \frac{312}{65.6} = 4.76$$

La *F* crítica se busca en la tabla correspondiente a la distribución *F*, para el nivel de significación correspondiente (en este caso, el 0.05), con 2 grados de libertad para el numerador (número de tratamientos menos 1) y 15 grados de libertad para el denominador (número de datos menos número de tratamientos). La *F* crítica es 3.68, ya que

$$P(F > 3.68 \mid 2{,}15) = 0.05$$

9. Tomar la decisión. Se rechaza la hipótesis nula si la *F* empírica es mayor que la *F* crítica. Como la *F* empírica, 4.76, es mayor que la *F* crítica 3.68, se rechaza la hipótesis nula y se concluye que cuando menos una de las medias no es igual a las otras.

   Se notó que no se empleó para el cálculo la suma de cuadrados total, se calculó para hacer evidente que esta *SCT* es igual a la suma de las otras variaciones, tal como se observó arriba; pero, aunque no se utilizó, puede servir para calcular otras sumas de cuadrados simplemente despejando en la ecuación que relaciona las 3 sumas de cuadrados o, también, calculando las 3, para asegurarse de que todos los cálculos están bien.

Una vez revisada la técnica, es importante hacer algunas consideraciones al respecto. La hipótesis nula implica que todas las medias son iguales y que, por lo mismo, las variaciones que se pueden dar son únicamente variaciones aleatorias. Si los 2 promedios de cuadrados se parecen (el *PCTr* y el *PCE*, el que se obtiene entre tratamientos y el debido a la aleatoriedad, respectivamente), eso quiere decir que podemos esperar que, efectivamente, la única variación que se presenta es la aleatoria, por eso no se rechaza la hipótesis nula; en cambio, si la variación entre tratamientos es mucho mayor que la variación aleatoria, sospechamos que sí existe una variación entre los diversos tratamientos y, por lo mismo, rechazamos la hipótesis nula y damos por buena la alternativa.

Para el director de producción de la fábrica de cuadros para bicicleta, la conclusión indica que no todos los turnos trabajan igual, sino que existen diferencias entre ellos. Se puede afirmar que su investigación apenas comienza, pues la diferencia entre las producciones medias indica (en este ejemplo en particular) que no todos los turnos son igualmente productivos y que se puede mejorar en alguno (o algunos) de ellos.

Se presenta a continuación otro ejemplo, ahora resuelto con ayuda de una hoja de cálculo de Excel.

Ejemplo 12.5

Se diseñaron 4 tipos diferentes de examen para evaluar el aprovechamiento en un curso de capacitación y, para probar si existen diferencias significativas en el diseño de los exámenes, se eligió un conjunto de 40 trabajadores en capacitación y se les asignó uno de los 4 exámenes al azar, los tratamientos. Los resultados se muestran en la tabla 12.1 y se incluyen los promedios de cada tratamiento.

Como los exámenes se asignaron aleatoriamente a los trabajadores, se pensaría que las diferencias entre los promedios de las calificaciones en los 4 subconjuntos (tratamientos) se deben a diferencias entre los propios exámenes, salvo diferencias aleatorias o errores normales de muestreo que en este caso pueden deberse a diferencias personales entre los examinados. La hipótesis nula sería aquí que no existe diferencia entre las medias de las calificaciones obtenidas a través de los 4 exámenes o, en símbolos:

$$H_0: \ _1 = \ _2 = \ _3 = \ _4$$

Tabla 12.1 Calificaciones de 40 trabajadores en capacitación, 10 por cada tipo de examen

	Exámenes			
	A	B	C	D
	71	84	79	92
	72	94	92	70
	80	77	73	74
	70	84	86	70
	85	96	82	74
	77	84	98	85
	79	86	98	70
	95	99	82	75
	63	96	91	62
	80	86	64	90
Promedio	77.2	88.6	84.5	76.2

La hipótesis alternativa plantea:

H_1: Cuando menos 2 de esas medias no son iguales.

Si las medias de las 4 poblaciones son iguales, es de esperar que las medias muestrales observadas sean similares y, por lo tanto, si se observan diferencias considerables en las medias muestrales se tendría evidencia para rechazar la hipótesis nula y aceptar la alternativa.

Además, si las varianzas entre los 4 grupos de calificaciones son las mismas (uno de los supuestos en los que se basa este análisis de varianza) entonces se puede pensar que, en realidad, las 40 observaciones proceden de una "gran" población, o población total, que abarca como subconjuntos a las 4 poblaciones en las que se dividieron las calificaciones, y que tiene una "gran" varianza, o varianza total.

Como se dijo antes, el análisis de varianza de una vía se basa en la relación que indica que la suma de cuadrados total, *SCT*, es igual a la suma de los cuadrados de tratamiento, *SCTr*, más la suma de cuadrados del error, *SCE*, o, en símbolos $SCT = SCTr + SCE$.

Para obtener estas 3 sumas de cuadrados, se requieren, en primer lugar, los promedios por columna (por tratamiento) y el promedio total. Estos promedios se pueden obtener fácilmente en Excel, con la función "Promedio". Si los datos están en las celdas A1 a D10, los promedios por columna se obtienen, por ejemplo, como =PROMEDIO(A1:A10), para la columna A y luego copiando esta función hasta la columna D. El promedio global se puede calcular promediando estos 4 promedios de columna o promediando todos los 40 datos. Se puede verificar fácilmente que el pro medio de los promedios de tratamientos (el promedio de los promedios por columna) es igual al promedio de todos los datos.

En la tabla 12.2 se resumen estos cálculos.

Tabla 12.2 Cálculos obtenidos por Excel

	Exámenes			
	A	B	C	D
Promedios	77.2	88.6	84.5	76.2

Y la media global sería:

$$\bar{X}_g = \frac{77.2 + 88.6 + 84.5 + 76.2}{4} = 81.625$$

Excel Se requiere ahora calcular la suma de cuadrados total, es decir, la suma de los cuadrados de las desviaciones de cada dato respecto a esta media global; de nuevo, en Excel basta con la función =(A1-81.625)^2 en, por ejemplo, la celda F1 y luego copiarla para cubrir todas las celdas hasta la I10. Esta nueva matriz de 10 por 4 contiene todos los cuadrados de las desviaciones; finalmente, sumando todos estos valores, =SUMA(F1:I10), se obtiene la suma de cuadrados total, que es:

$$SCT = 4\,139.375$$

Por su parte, la suma de cuadrados de tratamientos se obtiene elevando al cuadrado las diferencias entre cada uno de los promedios de columnas y el promedio global, para luego multiplicar cada una de estas diferencias al cuadrado por el tamaño de muestra y, finalmente, sumando todos estos productos.

Volviendo a los datos, que se resumieron en la tabla 12.2, las operaciones son como sigue:

$$SCTr = 10(77.2 - 81.625)^2 + 10(88.6 - 81.625)^2 + 10(84.5 - 81.625)^2 + 10(76.2 - 81.625)^2 = 195.81 + 486.51 + 82.66 + 294.31 = 1\,059.29$$

La suma de cuadrados del error, *SCE*, se calcula mediante la suma de cuadrados de la diferencia entre los datos de cada tratamiento (columna) y su correspondiente media. Esto con Excel es bastante sencillo. Si los datos están en las celdas A1 a D10 y luego se calculan los promedios por columna en las celdas A11 a D11, se tendría la matriz que se muestra a la derecha.

	A	B	C	D
1	71	84	79	92
2	72	94	92	70
3	80	77	73	74
4	70	84	86	70
5	85	96	82	74
6	77	84	98	85
7	79	86	98	70
8	95	99	82	75
9	63	96	91	62
10	80	86	64	90
11	77.2	88.6	84.5	76.2

Ahora, para calcular los cuadrados de las diferencias entre cada dato y su correspondiente media, se puede anotar la fórmula "=(A1-A$11)^2" colocada, por ejemplo, en la celda F1 da el cuadrado de la diferencia entre el primer dato del primer tratamiento.[1] Si ahora se copia esta fórmula, desde la celda F, sobre el mismo renglón 1 y hasta la celda I1 y luego se copian estos renglones F1 a I1 hasta el renglón 10 se tendría la matriz de 10 por 4 con todas estas diferencias al cuadrado. Finalmente, la suma de todas ellas, =Suma(F1:I10), da la suma de cuadrados del error, que es $SCE = 3\,080.10$.

[1] El signo $ que precede al 11 le indica a Excel que, aunque al copiar hacia abajo debe ir moviéndose de renglón, el dato sobre esa columna y el renglón 11 (es decir, la media de la columna) no debe moverse.

Una manera más sencilla de obtener esta *SCE* consiste en aprovechar la relación *SCT* = *SCTr* + *SCE* que, despejando, se convierte en: *SCE* = *SCT* – *SCTr*. Como estas 2 últimas sumas de cuadrados ya se calcularon, se tiene que *SCE* = 4 139.375 – 1 059.29 = 3 080.09, que es el mismo valor que se obtuvo con el procedimiento completo descrito antes (la pequeña diferencia se debe a redondeo).

Sin embargo, cuando sea necesario, se puede utilizar ese procedimiento completo para asegurarse de que las operaciones son correctas.

Ahora, para determinar el valor calculado del estadístico de prueba F, se calcula el promedio de cuadrados de tratamientos *PCTr* y el promedio de cuadrados del error, *PCE*. Ambos se obtienen dividiendo las correspondientes sumas de cuadrados entre sus grados de libertad. Los grados de libertad para la *SCTr* es el número de tratamientos menos 1, en tanto que los grados de libertad de la *SCE* es el número total de elementos (datos) menos el número de tratamientos. Con los datos del ejemplo se tiene que:

$$PCTr = \frac{SCTr}{gl_{Tr}} = \frac{1\,059.29}{4 - 1} = 353.10$$

$$PCE = \frac{SCE}{gl_{E}} = \frac{3\,080.09}{36} = 85.56$$

De donde la *F* empírica es:

$$F_e = \frac{PCTr}{PCE} = \frac{353.10}{85.56} = 4.13$$

Por su parte la *F* crítica, para un nivel de significación del 0.05 es 2.866, ya que

$$P(F > 2.866 \mid 3{,}36) = 0.05$$

Así, como el valor empírico de *F*, 4.13, es mayor que el valor crítico, 2.866, se rechaza la hipótesis nula y se concluye que al menos una de las medias de tratamiento no es igual a las otras.

En la sección siguiente se presenta un resumen del procedimiento para llevar a cabo análisis de varianza con el diseño completamente aleatorizado de una vía.

12.4 Procedimiento para el ANOVA con el diseño completamente aleatorizado de un factor

En estos casos se tienen 3 o más muestras, todas del mismo tamaño, y se tienen *n* observaciones en cada una de ellas, de manera que los datos conforman una matriz en la cual cada columna representa un tratamiento.

Se pueden representar sus elementos mediante un elemento general, *X*, que representa, precisamente, cada dato, con 2 subíndices, *i* y *j*, que representan el renglón y la columna que ocupan, respectivamente. En el cuadro 12.1 se hace una representación genérica para *n* elementos por cada muestra (el número de renglones) y *q* tratamientos o número de muestras.

Cuadro 12.1 Los datos de una matriz para ANOVA con *n* elementos por cada muestra (el número de renglones) y *q* tratamientos o número de muestras

$X_{1,1}$	$X_{1,2}$	…	$X_{1,q}$
$X_{2,1}$	$X_{2,2}$	…	$X_{2,q}$
$X_{i,1}$	$X_{i,j}$	…	$X_{3,q}$
…	…	…	
$X_{n,1}$	$X_{n,2}$	…	$X_{n,q}$

1. El primer paso, como en todos los casos de pruebas de hipótesis, consiste en plantear las hipótesis.
2. Para realizar la prueba de hipótesis sobre la igualdad de las medias de todos los tratamientos, en primer lugar se obtiene la media aritmética de los datos de cada muestra (tratamiento) y una media global comenzando con todos los datos en su conjunto; así, el primer paso consiste en obtener esas medias de tratamiento y, a partir de ellas, la media global (es fácil comprobar que esta media global es el promedio de las medias de todas las columnas). Se representa esto en el cuadro 12.2.

Cuadro 12.2 Los datos de una matriz para ANOVA con las medias por tratamiento (columna) y la media global

$X_{1,1}$	$X_{1,2}$	…	$X_{1,q}$
$X_{2,1}$	$X_{2,2}$	…	$X_{2,q}$
$X_{i,1}$	$X_{i,j}$	…	$X_{3,q}$
…	…	…	
$X_{n,1}$	$X_{n,2}$	…	$X_{n,q}$
$\bar{X}_{i,1}$	$\bar{X}_{i,2}$	…	$\bar{X}_{i,q}$

En donde

$$\bar{X}_{i,q} = \frac{\sum_{i=1}^{n} X_{i,q}}{n} \tag{12.2}$$

La media general o global es:

$$\bar{X}_{g} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{q} X_{i,q}}{nq} \tag{12.3}$$

3. El tercer paso del procedimiento consiste en calcular las diferentes sumas de cuadrados implicadas en la relación:

$$SCT = SCTr + SCE$$

La suma de cuadrados total se obtiene calculando, para cada elemento de la matriz, el cuadrado de la diferencia entre ese dato y la media global:

$$SCT = \sum_{i=1}^{n} \sum_{j=1}^{q} \left(X_{i,j} - \bar{X}_{g}\right)^2 \tag{12.4}$$

La suma de cuadrados de tratamientos se obtiene elevando al cuadrado las diferencias entre cada uno de los promedios de columnas (tratamientos) y el promedio global para, luego, multiplicar cada una de estas diferencias al cuadrado por el tamaño de muestra y, finalmente, sumando todos estos productos:

$$SCTr = n\sum_{i=1}^{n} \left(\bar{X}_{i,j} - \bar{X}_{g}\right)^2 \tag{12.5}$$

A su vez, la suma de cuadrados del error es la suma de cuadrados de la diferencia entre los datos de cada tratamiento (columna) y su correspondiente media:

$$SCE = \sum_{i=1}^{n} \sum_{j=1}^{q} \left(X_{i,j} - \bar{X}_{i}\right)^2 \tag{12.6}$$

4. El cuarto paso consiste en calcular los promedios de cuadrados de tratamientos y del error, los cuales se determinan como:

$$PCTr = \frac{SCTr}{gl_{Tr}} \tag{12.7}$$

Es decir, el promedio de cuadrados de tratamiento se obtiene dividiendo la suma de cuadrados de tratamiento que se obtuvo antes entre su correspondiente número de grados de libertad, que es igual al número de tratamientos menos 1:

$$gl_{Tr} = q - 1$$

Para luego:

$$PCE = \frac{SCE}{gl_E} \tag{12.8}$$

Este promedio de cuadrados del error se obtiene dividiendo la suma de cuadrados del error entre su número de grados de libertad, el cual es igual al número total de elementos en todas las muestras (nq), o sea número de renglones por número de columnas en la matriz de datos, menos el número de tratamientos: $gl_E = nq - q$.

5. El quinto paso consiste en determinar el valor empírico del estadístico de prueba F_e, el cual es simplemente el cociente entre el $PCTr$ y el PCE:

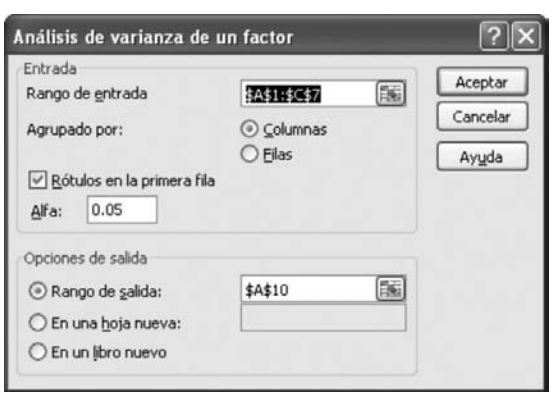
$$F_e = \frac{PCTr}{PCE} \tag{12.9}$$

6. El sexto paso consiste en determinar la F crítica para un nivel de significación, dado el cual se puede obtener de las tablas del apéndice o con la función Distr.F.Inv de Excel.
7. El séptimo paso, y final, consiste en aceptar la hipótesis nula cuando la F empírica es menor que la F crítica o en rechazarla en caso contrario, después de lo cual sólo resta interpretar este resultado en términos de las condiciones planteadas.

Del resumen anterior de los procedimientos se puede ver que el modelo para el análisis de varianza de un factor es:

$$SCT = SCTr + SCE$$

$$\sum_{i=1}^{n} \sum_{j=1}^{q} \left(X_{i,j} - \bar{X}_g\right)^2 = n\sum_{i=1}^{n} \left(\bar{X}_i - \bar{X}_g\right)^2 + \sum_{i=1}^{n} \sum_{j=1}^{q} \left(X_{i,j} - \bar{X}_i\right)^2 \tag{12.10}$$



12.5 Excel y ANOVA de un factor

En el complemento de Excel de análisis de datos se incluye una herramienta denominada "Análisis de varianza" de un factor que realiza todas las operaciones detalladas antes e ilustradas en los ejemplos 12.1 y 12.2.

Si se copian los datos del ejemplo 12.1 en una hoja de Excel y se colocan, incluyendo los títulos, en las celdas A1:C7 y después se ingresa a "Análisis de datos" y se selecciona "Análisis de varianza de un factor", se llega al cuadro de diálogo que se muestra a la izquierda.

En este cuadro ya se anotó el rango de celdas en donde se agrupan los datos por columnas, con rótulos en la primera fila y con nivel de significación, alfa, de 0.05 y con la solicitud de que se anoten los resultados a partir de la celda A10. Al marcar "Aceptar" se obtienen los siguientes resultados:

Análisis de varianza de un factor						
Resumen						
Grupos	**Cuenta**	**Suma**	**Promedio**	**Varianza**		
Matutino	6	822	137	55.6		
Vespertino	6	846	141	53.6		
Nocturno	6	762	127	87.6		

Análisis de varianza						
Origen de las variaciones	**Suma de cuadrados**	**Grados de libertad**	**Promedio de los cuadrados**	***F***	**Probabilidad**	**Valor crítico para *F***
Entre grupos	624	2	312	4.75609756	0.02513726	3.682320344
Dentro de los grupos	984	15	65.6			
Total	1 608	17				

Resume exactamente todos los cálculos realizados antes al resolver ese ejemplo 12.1, salvo 3 detalles: no se imprime la media global pero, por otro lado, se calculan las varianzas de cada factor y se anota la probabilidad de obtener el valor empírico de la *F*, que en este caso fue de 0.02513726.

Con estos datos se puede realizar la prueba de hipótesis mediante el método del estadístico de prueba al verificar, como se hizo antes, que la *F* empírica, 4.76, es mayor que la *F* crítica, 3.68.

Además, con el valor de la probabilidad de la *F* empírica, 0.025, se puede también realizar la prueba mediante el método de la *P*, y llegar a la misma conclusión de rechazar la hipótesis nula, ya que este valor de probabilidad es menor que el nivel de significación de 0.05.

Por supuesto, el ejemplo 12.2 se puede resolver de la misma sencilla manera, utilizando esta herramienta de Excel. Se invita al lector a hacer la comprobación correspondiente.

EJERCICIOS 12.5 ANOVA con diseño completamente aleatorizado de un factor

1. Un restaurante tiene 3 meseros, se desea averiguar si el tiempo que tardan ellos para tomar una orden es, en promedio, igual. Se desea un nivel de significación de 0.05. Los datos se dan a continuación, el tiempo es en minutos.

Mesero 1	Mesero 2	Mesero 3
4	7	4
3	6	3
5	9	4
3	11	5

a) Formule las hipótesis.
b) Realice el análisis de varianza.
c) Analice si se debe rechazar o no la hipótesis nula.
d) Sugiera qué investigación se puede hacer a continuación.

2. Los estudiantes de artes de una universidad enviaron a la dirección una carta de protesta pidiendo que les redujeran la cuota semestral correspondiente a servicios de cómputo, aduciendo que ellos utilizaban las instalaciones menos que los estudiantes de otras escuelas. Para decidir sobre la razonabilidad de su protesta, el rector mandó hacer un estudio sobre el número de horas por semestre que los estudiantes de diversas carreras utilizan este tipo de instalaciones. Los datos en horas por semestre se muestran a continuación.

Artes	Ciencias	Ingeniería	Administración
32	33	36	38
37	45	46	41
30	32	42	34
48	48	33	43
44	35	37	32
47	36	36	50

a) Formule las hipótesis.
b) Realice el análisis de varianza.
c) Tome la decisión de rechazar o no la hipótesis nula.
d) Sugiera al rector la contestación que podría dar a los estudiantes de artes.

3. Un inversionista se interesa en instalar un centro comercial en una ciudad que tiene últimamente un rápido crecimiento; él localizó 3 regiones en las afueras de la ciudad y desea saber si los ingresos de los habitantes cercanos a esas áreas son iguales. Para ello hizo un estudio socioeconómico de algunas familias seleccionadas al azar en la cercanía de cada corrida. La información en miles de pesos mensuales se da a continuación.

Región A	Región B	Región C
24 000	22 800	26 100
23 600	26 400	23 600
23 000	23 800	24 300
23 300	21 200	28 100
27 600	24 800	24 100
23 100	20 100	26 200

a) Formule las hipótesis.
b) Realice el análisis de varianza.
c) Analice si se debe rechazar o no la hipótesis nula.

4. Un gerente de distribución y almacenamiento de una empresa que fabrica productos perecederos estudia diversos equipos de refrigeración para determinar si todos ellos tienen un consumo igual de energía eléctrica. Midió el consumo de varios equipos de igual tamaño sometidos a tareas similares. Los resultados en Kw/hora por semana de trabajo se dan a continuación.

Marca A	Marca B	Marca C	Marca D
478	580	573	501
567	452	544	504
574	480	428	478
515	410	579	426
542	571	475	403
526	596	439	468
582	502	588	504

a) Formule las hipótesis.
b) Realice el análisis de varianza.
c) Analice si se debe rechazar o no la hipótesis nula.
d) ¿Considera usted que todos los equipos funcionan de manera igualmente eficiente? Si no es el caso, qué estudios adicionales le sugeriría al gerente de distribución y almacenamiento.

5. Una envasadora de aceite de oliva produce latas de 4 L de ese producto y tiene 4 máquinas que lo envasan. El gerente de producción desea saber si todas las máquinas llenan las latas con la misma cantidad de producto; para ello obtienen muestras aleatorias de las latas llenadas por los diferentes equipos. Los datos se listan a continuación.

Máquina 1	Máquina 2	Máquina 3	Máquina 4
4.04	3.98	4.02	3.94
4.02	4.02	3.98	3.98
4.05	4.02	4.03	4.00
4.00	4.01	3.99	
4.02	4.01	4.00	

a) Formule las hipótesis.
b) Realice el análisis de varianza con un alfa de 0.05.
c) Con base en su análisis indique qué concluye en relación con las hipótesis y sugiera qué hacer al gerente de producción.

12.6 Comparaciones múltiples entre pares de medias de tratamiento

En algunas ocasiones es suficiente para el tomador de decisiones saber que las medias son iguales o que no lo son para el trabajo posterior. En otros casos, si se rechaza la hipótesis nula, puede ser necesario profundizar más para saber cuál o cuáles medias difieren de las demás. En esta situación se deben comparar todos los pares posibles de medias. Se describe a continuación la forma en la que se realiza esto, utilizando el ejemplo de la fábrica de cuadros para bicicleta de la sección anterior.

EJEMPLO 12.6

Las medias de los 3 tratamientos del ejemplo son diferentes: $\overline{X}_1 = 137$ para el primer turno, $\overline{X}_1 = 141$ para el segundo turno y $\overline{X}_1 = 127$ para el tercero; lo importante es definir si esas diferencias son significativas para un determinado nivel de confianza o α. En este caso α de 0.05 permanecerá.

	Total	De tratamientos	Del error
Suma de cuadrados	1 608	624	984
Grados de libertad		2	15
Promedio de cuadrados		312	65.6

	1 matutino	2 vespertino	3 nocturno
n	6	6	6
$\overline{X}$	137	141	127

1. En primer lugar, se obtienen los valores absolutos de las diferencias entre todos los pares de medias posibles:

$$\left|\overline{X}_1 - \overline{X}_2\right| = \left|137 - 141\right| = 4$$
$$\left|\overline{X}_1 - \overline{X}_3\right| = \left|137 - 127\right| = 10$$
$$\left|\overline{X}_2 - \overline{X}_3\right| = \left|141 - 127\right| = 14$$

2. Se obtiene la diferencia significativa mínima (DSM) para cada caso, de la siguiente manera:

$$DSM = t\sqrt{PCE\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

En donde PCE es el promedio de cuadrados del error, (65.6).

n_1 y n_2 son el número de elementos del primer tratamiento y del segundo tratamiento (de los comparados en la dife-

rencia de medias), respectivamente; como todas las muestras son del mismo tamaño, n_1 y n_2 son, en este caso, siempre iguales a 6.

t es la t de student para un número de grados de libertad igual al número de datos menos el número de tratamientos ($gl = 18 - 3 = 15$) y teniendo en cuenta que se trata de una prueba de 2 colas. Aquí, con $\alpha = 0.05$ y 15 grados de libertad, el valor correspondiente de t es 2.131, pues:

$$P(-2.131 \geq t \geq 2.131 \mid gl = 15) = 0.05$$

Con estos datos, se tiene:

$$DSM = t\sqrt{PCE\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = 2.131\sqrt{65.6\left(\frac{1}{6} + \frac{1}{6}\right)}$$

$$= 2.131(4.676) = 9.965$$

Puesto que, como ya se mencionó, todas las muestras son del mismo tamaño, 6, la DSM es igual para todos los pares que deseamos comparar, en caso de muestras de diferente tamaño, cada par variará en consecuencia.

3. Se compara la diferencia absoluta de cada una de las medias apareadas con la diferencia significativa mínima; si es mayor que la DSM, se concluye que esas 2 medias son diferentes; si es menor, se concluye que no existe diferencia significativa. Desde luego, con el nivel de confianza definido en un principio.

Volviendo a las diferencias absolutas calculadas antes, se tiene que:

$$\left|\overline{X}_1 - \overline{X}_2\right| = \left|137 - 141\right| = 4$$
$$\left|\overline{X}_1 - \overline{X}_3\right| = \left|137 - 127\right| = 10$$
$$\left|\overline{X}_2 - \overline{X}_3\right| = \left|141 - 127\right| = 14$$

Sólo la primera diferencia es menor que la diferencia significativa mínima y, por ello, la diferencia entre la media del turno matutino y el vespertino no es significativa, por lo que se concluye que esas 2 medias son iguales.

Por otra parte, como las diferencias absolutas entre las medias del turno matutino con el turno nocturno y entre las de los turnos vespertino y nocturno son superiores a esa DSM se concluye, entonces, que las medias de esos 2 pares de turno sí son diferentes.

En otras palabras, el director de producción de la fábrica puede concluir que no existe diferencia entre el turno matutino y el vespertino, pero sí entre ellos y el turno de noche.

EJERCICIOS 12.6 Pruebas entre pares de medias de tratamientos

Tomando como base los problemas del 1 al 5 de la sección anterior, en caso de que se rechazara la hipótesis nula, realice las pruebas para diferencias apareadas y exponga conclusiones completas con base en las circunstancias del caso.

12.7 Análisis de varianza de dos factores

En el análisis de varianza de un factor se supone que la variación total puede derivarse únicamente de dos fuentes: la variación entre tratamientos y la aleatoria. En otras circunstancias puede sospecharse otra fuente de variación. El siguiente ejemplo, el mismo que se presentó en el ejemplo 12.2, servirá para explicarlo.

EJEMPLO 12.7

Una pequeña empresa que elabora botanas (papas fritas, cacahuates, etc.) tiene 4 rutas para el reparto de sus productos dentro de la ciudad. Para cubrir estas rutas tiene 5 conductores, el conductor complementario sirve para cubrir ausencias por vacaciones o enfermedad de los otros 4 y se le ocupa para otros menesteres si no se le requiere como conductor. El puesto de conductor complementario se rota entre todos los choferes sobre una base mensual. Es importante, tanto desde el punto de vista de eficiencia, como de justicia, que todas las rutas tengan la misma duración. También es importante que todos los conductores sean igualmente eficientes en todas las rutas, para que los conductores puedan sustituirse el uno al otro sin mermas de ninguna clase. Por ello, la gerencia de distribución implementó un programa de capacitación para que los conductores se familiaricen con todas las rutas; al término de esta capacitación hubo recorridos de prueba con el objeto de probar, por una parte, si todas las rutas se recorrían en el mismo tiempo y, por otra, si los conductores eran igualmente eficientes en todas. La gerencia de distribución desea saber si el tiempo medio de las 4 rutas es igual y si la eficiencia media de todos los conductores en las rutas es igual. La información de los recorridos y los conductores se da a continuación.

Conductor/ruta	Ruta A	Ruta B	Ruta C	Ruta D
Antúnez	224	227	237	248
Becerra	242	235	262	250

(*continúa*)

(*continuación*)

Conductor/ruta	Ruta A	Ruta B	Ruta C	Ruta D
Cervantes	225	240	235	261
Domínguez	232	253	259	255
Escamilla	232	245	257	261

NOTA El tiempo se mide en minutos.

Se trata de identificar la igualdad o no igualdad, tanto de los tiempos de recorrido entre las diferentes rutas como de la eficiencia de los conductores (los dos "factores"). Nótese que la variable de las rutas es lo que antes equivalió al "tratamiento" y era el único factor que se consideró en el ANOVA de un factor: es la variable que se divide en las 4 columnas: las 4 rutas; a la otra variable, los conductores, se le suele llamar "bloques" y es la que se dividió en los 5 renglones y constituye el segundo factor que se considera. En estos términos, se intenta identificar la igualdad o no igualdad tanto de tratamientos como de bloques, los dos factores. Es decir, tratamos de discernir si todas las rutas tienen medias iguales y, por otro lado, si todos los conductores tienen medias iguales. Identificaremos las medias de la siguiente manera:

En relación con las rutas:

μ_A = Tiempo medio de la ruta A.
μ_B = Tiempo medio de la ruta B.
μ_C = Tiempo medio de la ruta C.
μ_D = Tiempo medio de la ruta D.

En relación con los conductores:

μ_1 = Tiempo medio del conductor Antúnez.
μ_2 = Tiempo medio del conductor Becerra.
μ_3 = Tiempo medio del conductor Cervantes.
μ_4 = Tiempo medio del conductor Domínguez.
μ_5 = Tiempo medio del conductor Escamilla.

Se identifican 3 variables: las rutas, que ya conocemos como factor 1, y los choferes, a la que llamaremos factor 2. Ambas son variables independientes. También tenemos los tiempos de recorrido que es la variable dependiente o de respuesta. Asimismo, tenemos 3 fuentes de variación, que integran la variación total: las diferencias entre tratamientos (las 4 diferentes rutas, el factor 1) y la diferencia aleatoria o del error ya se vio en la sección de ANOVA de 1 factor; a éstas, agregaremos la posible variación entre los repartidores (el factor 2).[2] Entonces, se tendrán ahora 4 sumas de cuadrados:

SCT. La suma de cuadrados de la variación total.
SCTr. La suma de cuadrados de la variación entre tratamientos.
SCE. La suma de cuadrados de la variación del error, o aleatoria.
SCB. La suma de cuadrados de la variación entre bloques.

El modelo se convierte en:

$$SCT = SCTr + SCB + SCE$$

[2] Diversos autores utilizan distintas nomenclaturas. Algunos de ellos en lugar de llamar tratamientos y bloques a las variables independientes, les llaman *tratamientos y vías.*

Se espera que al tener una nueva variable que explique la variación total (la variable de bloque) se reducirá la variación aleatoria. Al reducirse la variación aleatoria, que es el denominador de la F empírica, la razón como un todo se incrementará y si efectivamente existen variaciones no aleatorias (debidas a los tratamientos o a los bloques), éstas serán más fáciles de identificar y tendremos mejores elementos para rechazar la hipótesis nula. Para explicar la técnica, nos referiremos al ejemplo de la fábrica de botanas ya planteado: la mayoría de las actividades es muy similar al caso de una vía. De encontrarse diferencias se explicarán con detalle.

1. Plantear las hipótesis. Dada la explicación anterior, se tienen 2 juegos de hipótesis: las relativas a los tratamientos (rutas) y las relativas a los bloques (conductores).

   Para las rutas:

   H_0: $\mu_A = \mu_B = \mu_C = \mu_D$
   H_1: por lo menos una de las medias poblacionales de los tratamientos no es igual a las otras.

   Para los conductores:

   H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$
   H_1: por lo menos una de las medias poblacionales de los bloques (conductores) no es igual a las otras.

2. Obtener las medias: la media de cada tratamiento, la media de cada bloque y la media general de todos los datos:

Medias de los tratamientos

$$\bar{X}_A = \frac{(224 + 242 + 225 + 232 + 232)}{5} = 231$$

$$\bar{X}_B = \frac{(227 + 235 + 240 + 253 + 245)}{5} = 240$$

$$\bar{X}_C = \frac{(237 + 262 + 235 + 259 + 257)}{5} = 250$$

$$\bar{X}_D = \frac{(248 + 250 + 261 + 255 + 261)}{5} = 255$$

Medias de los bloques

$$\bar{X}_1 = \frac{(224 + 227 + 237 + 248)}{4} = 234.00$$

$$\bar{X}_2 = \frac{(242 + 235 + 262 + 250)}{4} = 247.25$$

$$\bar{X}_3 = \frac{(225 + 240 + 235 + 261)}{4} = 240.25$$

$$\bar{X}_4 = \frac{(232 + 253 + 259 + 255)}{4} = 249.75$$

$$\bar{X}_5 = \frac{(232 + 245 + 257 + 261)}{4} = 248.75$$

En una hoja de cálculo:

	A	B	C	D	Medias de bloques
1	224	227	237	248	234
2	242	235	262	250	247.25
3	225	240	235	261	240.25
4	232	253	259	255	249.75
5	232	245	257	261	248.75
Medias de tratamientos	231	240	250	255	

Y, la media general:

$$\bar{X}_G = \frac{(224 + \cdots + 232 + 227 + \cdots + 245 + 237 + \cdots + 257 + 248 + \cdots + 261)}{20} = 244$$

Nótese que esta media general se puede obtener con el promedio de las medias de tratamiento o, alternativamente, como el promedio de las medias de bloque.

3. Obtener las diferentes sumas de cuadrados:

La suma de cuadrados total

$$SCT = (224 - 244)^2 + \cdots + (232 - 244)^2 + (227 - 244)^2 + \cdots + (245 - 244)^2 + (237 - 244)^2 + \cdots + (257 - 244)^2 + (248 - 244)^2 + \cdots + (261 - 244)^2 = 3\,120$$

En un cuadro de Excel, que contiene todas las diferencias al cuadrado:

400	289	49	16
4	81	324	36
361	16	81	289
144	81	225	121
144	1	169	289

La suma de todas ellas es, por supuesto, 3 120.

La suma de cuadrados de tratamientos:

$$SCTr = 5(231 - 244)^2 + 5(240 - 244)^2 + 5(250 - 244)^2 + 5(255 - 244)^2 = 1\,710$$

La suma de cuadrados de bloques

Para los bloques, se multiplica, como se hizo en el caso de los tratamientos, cada diferencia de cuadrados por el número de elementos de cada bloque.

$$SCB = 4(234 - 244)^2 + 4(247.25 - 244)^2 + 4(240.25 - 244)^2 + 4(249.75 - 244)^2 + 4(248.75 - 244)^2 = 721$$

La suma de cuadrados del error

En este caso se debe hacer lo siguiente: a cada dato se le resta la media del tratamiento que le corresponde y la media del bloque que le corresponde y se le suma la media total; se eleva al cuadrado el resultado de esta operación y, finalmente, se suman todos los cuadrados.

$$SCE = (224 - 231 - 234 + 244)^2 + \cdots + (255 - 225 - 248.75 + 244)^2 = 689$$

En una hoja de cálculo de Excel:

9	9	9	9
60.0625	68.0625	76.5625	68.0625
5.0625	14.0625	126.5625	95.0625
22.5625	52.5625	10.5625	33.0625
14.0625	0.0625	5.0625	1.5625

La suma de todos estos cuadrados es 689.

Excel El lector se puede percatar de que ese último cálculo es bastante laborioso, por lo que la manera sencilla de hacerlo es mediante un cuadro de Excel como el anterior, en donde la entrada de la celda de arriba a la izquierda es: =(B2-$F2-B$7+244)^2, si los datos originales están en las celdas B2:E6. También, como se vio antes, es posible obtener esta *SCE* mediante la diferencia entre la *SCT* – *SCTr* – *SCB* según se desprende del modelo, pero recordando que hacer el cálculo completo ayuda a asegurar la corrección de los resultados.

Resumiendo:

$SCT = 3\,120$
$SCTr = 1\,710$
$SCB = 721$
$SCE = 689$

Se verifica que $SCT = SCTr + SCB + SCE = 1\,710 + 721 + 689 = 3\,120$.

4. Obtener los promedios de cuadrados, tanto de tratamientos *(PCTr)*, como de bloques *(PCB)*, dividiendo en ambos casos entre el número de grados de libertad (4 − 1 en el caso de los tratamientos y 5 − 1 en el caso de los bloques):

$$PCTr = \frac{1\,710}{3} = 570$$

$$PCB = \frac{721}{4} = 180.25$$

Y, además, obtener el promedio de los cuadrados del error. Se obtiene dividiendo la suma de cuadrados del error entre el número de tratamientos menos 1 multiplicado por el número de bloques menos 1; es decir (4 − 1) (5 − 1) = 12, que es también, el número de grados de libertad en este caso:

$$PCE = \frac{689}{12} = 57.42$$

5. Obtener los valores de *F*. Tanto los empíricos para tratamientos y bloques, como los críticos de la tabla, con un α, un nivel de significación, de 0.05.

Para tratamientos:

$$F_e = \frac{PCTr}{PCE} = \frac{570}{57.42} = 9.92$$

La *F* crítica es 3.49, ya que, de la tabla de la distribución *F*, 3 grados de libertad para el numerador (número de tratamientos menos 1) y 12 para el denominador (número de tratamientos menos 1 por número de bloques menos 1), o en símbolos:

$$P(F \geq 3.49 \mid gl = 3, gl = 12)$$

Para bloques:

$$F_e = \frac{PCB}{PCE} = \frac{180.25}{57.42} = 3.14$$

La *F* crítica es 3.26, ya que, de la tabla de la distribución *F*, 4 grados de libertad para el numerador (número de bloques menos 1) y 12 para el denominador (número de tratamientos menos 1 por número de bloques menos 1), o en símbolos:

$$P(F \geq 3.26 \mid gl = 4, gl = 12)$$

6. Tomar la decisión. Se rechaza la hipótesis nula si la *F* empírica es mayor que la *F* crítica. En el ejemplo, tanto para el caso de los tratamientos como en el caso de los bloques, la $F_e > F_{cr}$, por lo que en ambos casos se debe rechazar la hipótesis nula y dar por buena la hipótesis alterna; se concluye entonces que las medias de los tiempos de reparto en las rutas no son todas iguales y, también, que las medias de los tiempos de reparto de todos los conductores tampoco son todas iguales.

Y, siendo así, si el objetivo de la gerencia de distribución es tener un desempeño uniforme de rutas y repartidores, debe tomar medidas tanto para igualar las rutas de reparto, como para capacitar mejor a los conductores en todas las rutas.

En las pruebas ANOVA de 2 vías se tienen 4 posibilidades:

1. No se rechaza ninguna de las hipótesis nulas.
2. Se rechaza la hipótesis de tratamientos iguales pero no la de bloques iguales.
3. Se rechaza la hipótesis de bloques iguales pero no la de tratamientos iguales.
4. Se rechazan ambas.

Las 4 situaciones tienen consecuencias desde el punto de vista del administrador. Éstas se examinan brevemente a continuación desde el punto de vista del problema de distribución que se acaba de resolver.

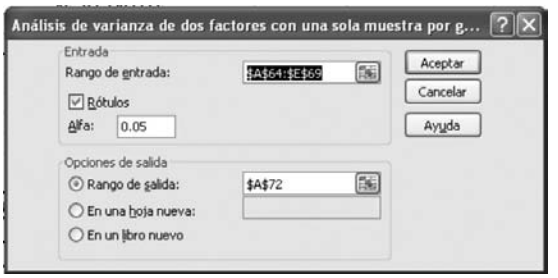
1. Si no se rechaza ninguna de las hipótesis, se puede afirmar que la gerencia logró, en este caso, su objetivo de equilibrio de rutas y de conductores, por lo que la situación en general solamente requiere de seguimiento.
2. Si se rechaza la hipótesis de tratamientos iguales pero no la de bloques iguales, significa que los operadores son igualmente eficientes en todas las rutas, pero que estas últimas no son iguales, por lo que la dirección deberá de nivelarlas. La actividad de la dirección de distribución será la de igualar las rutas.
3. Si se rechaza la hipótesis de bloques iguales pero no la de tratamientos iguales, en este caso las rutas sí están bien diseñadas, pero se debe tomar acción correctiva en relación con la capacidad de los conductores de manejarse en ellas.
4. Si se rechazan ambas, la dirección debe tomar medidas correctivas tanto en la nivelación de rutas como en la capacitación de conductores tal como ya se explicó.

Se debe recordar que la información del ANOVA nos indica la situación pero no sus causas, por lo que la dirección deberá identificarlas para ejecutar acciones correctivas efectivas.

12.8 Excel y ANOVA de dos factores

En el menú Complementos de Excel, en el submenú "Análisis de datos" se incluye una herramienta denominada "Análisis de varianza de dos factores con una sola muestra por grupo" que realiza todas las operaciones detalladas antes e ilustradas en el ejemplo 12.7. Nótese, de nuevo, que a lo que en este texto se denomina "2 vías", Excel lo llama "dos factores" (si ya está cargada esta herramienta, la puede encontrar en el menú Herramientas).

Si se copian los datos de ese ejemplo 12.7 en una hoja de Excel y se colocan, incluyendo los títulos, en las celdas A64:E69 y, después, se ingresa a *Análisis de datos* y se selecciona ese *Análisis de varianza de un factor*, se llega al cuadro de diálogo que se muestra a la derecha.



En este cuadro ya se anotó el rango de celdas en donde están los datos, con rótulos tanto de renglones como de columnas y con nivel de significación, alfa (α), de 0.05 y con la solicitud de que se anoten los resultados a partir de la celda A72. Al marcar "Aceptar", se obtienen los siguientes resultados:

Análisis de varianza de dos factores con una sola muestra por grupo						
Resumen	**Cuenta**	**Suma**	**Promedio**	**Varianza**		
Antúnez	4	936	234	118		
Becerra	4	989	247.25	134.25		
Cervantes	4	961	240.25	230.25		
Domínguez	4	999	249.75	146.25		
Escamilla	4	995	248.75	170.916667		
Ruta A	5	1 155	231	52		
Ruta B	5	1 200	240	97		
Ruta C	5	1 250	250	167		
Ruta D	5	1 275	255	36.5		
Análisis de varianza						
Origen de las variaciones	**Suma de cuadrados**	**Grados de libertad**	**Promedio de los cuadrados**	***F***	**Probabilidad**	**Valor crítico para *F***
Filas	721	4	180.25	3.13933237	0.05538481	3.259166727
Columnas	1 710	3	570	9.92743106	0.00142859	3.490294821
Error	689	12	57.4166667			
Total	3 120	19				

Estos resultados resumen exactamente todos los cálculos realizados antes al resolver el ejemplo 12.7, salvo los 3 detalles que se mencionaron en la herramienta de ANOVA para un factor (1 vía): no se imprime la media global pero, por otro lado, se calculan las varianzas de cada factor y se anota la probabilidad de obtener el valor empírico de las 2 *F*, que en este caso fueron de 3.259166727, para el factor o vía de los vendedores (renglones) y de 3.490294821 para el factor o vía de las zonas (columnas).

Al igual que se hizo con la otra herramienta de Excel, con estos datos se puede realizar la prueba de hipótesis mediante el método del estadístico de prueba y verificar que se llega a las mismas conclusiones.

EJERCICIOS 12.8 ANOVA de dos factores

1. Una empresa de taxis da servicio de transporte entre los 2 principales aeropuertos de la ciudad. De acuerdo con la experiencia, eligieron 3 rutas para hacer el recorrido y se seleccionaron 4 conductores al azar para hacer recorridos de prueba a la misma hora. ¿Se puede afirmar que las 3 rutas requieren el mismo tiempo y que los conductores

trabajan igualmente bien todas las rutas? Los datos de los recorridos en minutos se encuentran en la siguiente tabla. Use un nivel de significación de 0.01.

Conductores/rutas	1	2	3
A	73	70	66
B	76	71	67
C	68	72	73
D	77	72	75

2. Una empresa que asegura automóviles tiene 2 agencias en la ciudad para atender llamados de siniestros. En el cuadro siguiente se muestran los datos de los llamados atendidos por las 2 agencias y para los diferentes días de la semana. ¿Se puede decir que hay diferencia entre los llamados atendidos por las 2 agencias o por el día de la semana en que ocurren? Use un nivel de significación de 0.05.

	Agencia 1	Agencia 2
Lunes	52	42
Martes	47	51
Miércoles	54	53
Jueves	45	49
Viernes	50	57
Sábado	53	52
Domingo	47	45

3. Un proveedor de servicios por Internet tiene un centro de llamadas para atender las peticiones de auxilio técnico de sus clientes, este centro trabaja las 24 horas en 3 turnos y los operadores rotan turnos periódicamente. La gerencia está interesada en saber si el tiempo de respuesta a los clientes, desde que entra la llamada hasta que es contestada, es igual en los diferentes turnos y para los distintos empleados. ¿Qué puede informarle a la gerencia con un nivel de significación de 0.01?

	Turno		
Empleado	Matutino	Vespertino	Nocturno
Arteaga	62	50	71
Gómez	67	53	66
González	57	48	60
Martínez	61	59	55
Villegas	56	51	55

4. Una revista especializada en automóviles hace pruebas de eficiencia en el consumo de combustible de los modelos compactos de 3 fabricantes de automóviles. Hace las pruebas en 3 tipos de terreno: ciudad, terreno montañoso y terreno llano con poco tráfico. ¿Consideraría que hay evidencia de diferencia en el consumo de combustible de los carros y en los diferentes tipos de terreno? La información se expresa en km/L, use un nivel de confianza de 0.05.

	Fabricante 1	Fabricante 2	Fabricante 3
Ciudad	14	12.5	13.1
Montaña	15.3	14.5	14.2
Plano	16.1	15.6	16

5. Con el objeto de buscar los puntos más favorables para situar nuevas estaciones, una empresa operadora de estaciones de gasolina lleva a cabo un estudio para averiguar si la localización de la gasolinera y el número de competidores en el radio de 1 km influyen en el número de vehículos atendidos por hora, en el horario de servicio de 7 a.m. a 9 p.m. ¿Existe diferencia en el promedio de vehículos atendidos por hora entre las diferentes localizaciones?, ¿el número de competidores en las proximidades influye en el número de vehículos atendidos por hora? Realice un análisis de varianza de dos factores para responder a estas preguntas. Utilice un α de 0.01 con los datos que se presentan a continuación:

Localización	Cero competidores	Un competidor	Dos competidores	Tres o más competidores
En carretera	39	36	53	42
En los suburbios	26	32	47	46
En el centro de las ciudades	28	22	27	29

12.9 Análisis de varianza de dos factores con interacción

En el análisis de varianza de una vía, tal como ya se comentó, se supone que existían 2 fuentes de posible variación: la variación entre tratamientos y la aleatoria o de error; en el de dos vías se amplió la perspectiva para contemplar una nueva posible fuente de variación y se suponen 3 fuentes para la misma: la que se da entre tratamientos, la que se produce entre bloques y la aleatoria. En este último apartado se desarrollará el ANOVA con 1 fuente más de variación: la que se puede producir por la interacción entre los tratamientos y los bloques, al igual que se agrupa la nomenclatura de tratamientos y bloques, para llamar

a las 2 variables independientes "factores" con el propósito de facilitarla, ya que ahora se hablará de una fuente adicional de variación (de sumas de cuadrados): la de la interacción. Para explicar esta técnica de ANOVA de dos factores con interacción se recurrirá al siguiente ejemplo.

Ejemplo 12.8

Una empresa que fabrica y vende purificadores de agua para el comercio y la industria tiene 3 zonas de ventas: centro, norte y sur. Dado el hecho de que la labor de venta tiene aspectos tanto técnicos como financieros, la empresa tiene 3 vendedores: un químico, un licenciado en administración y un ingeniero mecánico. La gerencia de la empresa está interesada en saber si las 3 zonas tienen un potencial equivalente, si los 3 vendedores tienen igual capacidad y si es indistinto el trabajo de los vendedores en cualquier zona o existen diferencias. Esta ampliación de nuestra perspectiva requiere un poco más de información, es necesario contar con al menos 2 mediciones para cada combinación vendedor-zona; es decir, se tiene que replicar la observación. Si se tienen 2 observaciones para cada combinación vendedor-zona, se dice que se tienen 2 réplicas; si se tienen 3 observaciones, 3 réplicas y así sucesivamente. Las variables independientes son los factores: *zonas* y *vendedores*, la variable dependiente es el volumen de ventas. La información pertinente se da en la siguiente tabla, se verá que hay 2 datos para cada combinación zona-vendedor. Para esta técnica es indispensable tener por lo menos 2 observaciones para cada combinación de factores y el número de observaciones para cada una de estas combinaciones debe ser igual. Son 2 observaciones para cada una en nuestro ejemplo, pero pueden ser 3 o más, siempre que sea el mismo número para todas.

	Ventas mensuales en miles de pesos		
Profesión del vendedor	**Zona centro**	**Zona sur**	**Zona norte**
Químico	506	528	513
	512	534	495
Licenciado en administración	529	496	508
	525	498	500
Ingeniero mecánico	500	512	528
	518	504	520

Con esta información y la técnica de análisis de varianza de 2 factores con interacción, determine si todas las zonas son iguales, si todos los vendedores son igualmente eficaces y si no existen interacciones que hagan que algún vendedor trabaje mejor en alguna zona. Identificaremos las medias como sigue:

Para las zonas:

μ_c = Ventas promedio de la zona centro.
μ_s = Ventas promedio de la zona sur.
μ_n = Ventas promedio de la zona norte.

Para los vendedores:

μ_q = Ventas promedio del químico.
μ_a = Ventas promedio del licenciado en administración.
μ_i = Ventas promedio del ingeniero mecánico.

Se identifican 3 variables: las zonas, que se manejaron como *variable de tratamiento* y que ahora será uno de los factores: el factor A; la profesión de los vendedores, a la que se llamó *variable de bloque* y que ahora será el factor B. Ambas son variables independientes. También se tienen los volúmenes de venta que son la *variable dependiente* o *de respuesta*. Asimismo, tenemos 4 fuentes de variación que integran la variación total: las diferencias debidas al factor A (las debidas a las 3 diferentes zonas), las diferencias debidas al factor B (las debidas a los diferentes vendedores), las debidas a la interacción entre el factor A y el factor B y, finalmente, las debidas a las variaciones aleatorias, o de error, a la suma de cuadrados de la variación total se le sigue llamando *SCT*; se identifica como *SCFA* a la suma de cuadrados de la variación entre los elementos del factor A; *SCFB* a la suma de cuadrados de las variaciones debidas a los elementos del factor B; *SCAB* a la suma de cuadrados de las variaciones debidas a la interacción entre los 2 factores y *SCE* a la suma de los cuadrados del error.

Se tiene, entonces, que el modelo que identifica la relación entre la variación total y las variaciones parciales es:

$$SCT = SCFA + SCFB + SCAB + SCE$$

Se espera que, al tomar en cuenta la interacción entre los factores, se reducirá la variación aleatoria, al reducirse ésta, que es el denominador de la F empírica (la F_e del ejemplo anterior); la razón como un todo se incrementará, y si efectivamente existen variaciones no aleatorias (debidas a los factores y a la interacción entre ellos), éstas serán más fáciles de identificar y se tendrían mejores elementos para rechazar la hipótesis nula. En seguida, la solución del ejemplo de la empresa de purificadores.

Solución: La mayoría de las actividades es muy similar a los casos de 1 y de 2 vías, tal como a continuación se indica:

1. Plantear las hipótesis. Ahora se tienen 3 juegos de hipótesis, las relativas al factor A (zonas), las relativas al factor B (profesión de los vendedores) y las relativas a la interacción entre ambas.

   Para las zonas:

   H_0: $\mu_c = \mu_s = \mu_n$
   H_1: por lo menos una de las medias poblacionales del factor A (las zonas) no es igual a las otras.

   Para los vendedores:

   H_0: $\mu_q = \mu_a = \mu_i$
   H_1: por lo menos una de las medias poblacionales del factor B (los vendedores) no es igual a las otras.

   Para la interacción:

   H_0: no existe interacción entre la zona y el vendedor.
   H_1: sí existe interacción entre zona y vendedor.

Este último conjunto de hipótesis requiere una breve explicación. H_0 indica que suponemos que todos los vendedores se desempeñan igual en todas las zonas. H_1 indica que sí existe una diferencia y que diversos vendedores se desempeñarán de distinta manera en diferentes zonas; es decir, que en alguna zona venderán mejor que en las otras.

2. Obtener las medias. La media de los factores A y B, la de cada combinación de factores (vendedor-zona) y la general de todos los datos en una hoja de Excel, con los promedios por columna y por renglón, es decir, promedios por zona y promedios por tipo de vendedor:

	Ventas mensuales en miles de pesos			
Profesión del vendedor	**Zona centro**	**Zona sur**	**Zona norte**	**Promedio**
Químico	506	528	513	
	512	534	495	514.67
Licenciado en Administración	529	496	508	
	525	498	500	509.33
Ingeniero mecánico	500	512	528	
	518	504	520	513.67
Promedio	515	512	510.67	512.56

Para el factor A (zonas):

$$\overline{X}_c = 515$$
$$\overline{X}_s = 512$$
$$\overline{X}_n = 510.67$$

Para el factor B (profesiones/vendedores):

$$\overline{X}_q = 514.67$$
$$\overline{X}_a = 509.33$$
$$\overline{X}_i = 513.67$$

Para las combinaciones de muestras vendedor-zona (los promedios de cada par de observaciones por vendedor), en otra tabla de Excel.

	Ventas mensuales en miles de pesos					
Profesión	**Centro**	**Sur**	**Norte**	**Medias de combinaciones**		
Químico	506	528	513			
	512	534	495	509	531	504
Licenciado en Adminis-tración	529	496	508			
	525	498	500	527	497	504
Ingeniero mecánico	500	512	528			
	518	504	520	509	508	524

$$\overline{X}_{c,q} = 509$$
$$\overline{X}_{c,a} = 527$$
$$\overline{X}_{c,i} = 509$$
$$\overline{X}_{s,q} = 531$$
$$\overline{X}_{s,a} = 497$$
$$\overline{X}_{s,i} = 508$$
$$\overline{X}_{n,q} = 504$$
$$\overline{X}_{n,a} = 504$$
$$\overline{X}_{n,i} = 524$$

La media general que se puede obtener mediante el promedio del total de los datos o como el promedio de los promedios de renglón o de columna pero que, en última instancia, es prácticamente igual de sencillo de obtener con Excel con cualquiera de los 3 procedimientos

$$X_g = 512.56$$

3. Obtener las diferentes sumas de cuadrados, de nuevo con Excel.

La suma de cuadrados total es la suma de los cuadrados de la diferencia entre cada observación y la media global que, sumadas, arrojan una $SCT = 2\,694.44$.

43.03	238.39	0.19
0.31	459.67	308.35
270.27	274.23	20.79
154.75	211.99	157.75
157.75	0.31	238.39
29.59	73.27	55.35

La suma de cuadrados de las diferencias en el factor A: la suma de los cuadrados de la diferencia entre las medias de cada columna y el promedio general, multiplicadas por el número de observaciones (6).

Para el factor A:

$$SCFA = 6(515 - 512.56)^2 + 6(512 - 512.56)^2 + 6(510.67 - 512.56)^2 = 59.04$$
$$SCFA = 59.04$$

Para el factor B:

$$SCFB = 6(514.67 - 512.56)^2 + 6(509.33 - 512.56)^2 + 6(513.67 - 512.56)^2 = 96.7$$
$$SCFB = 96.7$$

Para calcular todos los cuadrados de las interacciones, se toma el promedio de la combinación vendedor-zona, se le resta la media de la zona y la media del vendedor y se le suma la media general. Este dato se eleva al cuadrado y se multiplica por el número de elementos de cada combinación (es decir, el número de réplicas, en nuestro caso 2). En la tabla siguiente se resumen los promedios de las combinaciones de muestras vendedor-zona (los promedios de cada par de observaciones por vendedor) que se presentaron antes, junto con los promedios por zona y por tipo de vendedor, para facilitar la visualización de las operaciones.

Profesión	Centro	Sur	Norte	Promedio por tipo de vendedor	Medias de combinaciones		
Químico	506	528	513				
	512	534	495	514.67	509	531	504
Lic. en Adm.	529	496	508				
	525	498	500	509.33	527	497	504
Ing. Mec.	500	512	528				
	518	504	520	513.67	509	508	524
Promedio por zona	515	512	510.67				

$$SCAB = 2(509 - 515 - 514.67 + 512.56)^2 + 2(527 - 515 - 509.33 + 512.56)^2 + \cdots + 2(524 - 510.67 - 513.67 + 512.56)^2$$

En el cuadro siguiente se presentan los resultados para cada una de las nueve combinaciones:

131.44	570.77	153.94
463.70	277.22	23.67
101.01	52.16	298.98

La suma de todos estos cuadrados:

$$SCAB = 2\,072.89$$

En el caso de la suma de cuadrados del error, el procedimiento consiste en elevar al cuadrado la diferencia entre cada dato y la media correspondiente a la combinación de los 2 factores A y B. Con referencia de nuevo a la tabla anterior, al elemento del extremo superior izquierdo, 506, se le resta la media de la combinación de factores, 509, y se eleva al cuadrado esta diferencia; otro ejemplo, al valor 500 correspondiente al segundo dato de la combinación de la zona norte (licenciado en administración), se le resta la media de esa combinación, el 504 que aparece en la parte derecha de la tabla y se eleva esta diferencia al cuadrado. Se puede resumir el conjunto de estas operaciones de la siguiente manera:

$$SCE = (506 - 509)^2 + \cdots + (520 - 524)^2 = 466$$

Al realizar estas operaciones con todos los elementos de la tabla original se llega al siguiente conjunto de cuadrados de diferencias, cuya suma es precisamente la suma de cuadrados del error que se anota arriba, 466.

9	9	81
9	9	81
4	1	16
4	1	16
81	16	16
81	16	16

Para verificar que los cálculos son correctos, se resumen en la ecuación que los relaciona:

$$SCT = SCFA + SCFB + SCAB + SCE$$
$$2\,694.44 = 59.04 + 96.7 + 2\,072.89 + 466$$

En realidad, la suma de esas 4 sumas de cuadrados es 2 694.63 y en la cual la ligera diferencia contra el 2 694.44 que se obtuvo por la suma de cuadrados total se debe a errores de redondeo.

4. Se obtienen ahora los promedios de cuadrados de los factores y de su interacción, con sus respectivos grados de libertad:

   Grados de libertad del factor A = número de factores A menos 1 = 3 – 1 = 2.
   Grados de libertad del factor B = número de factores B menos 1 = 3 – 1 = 2.
   Grados de libertad de la interacción entre los 2 factores = número de factores A menos 1, multiplicado por el número de factores B menos 1 = (3 – 1)(3 – 1) = 4.
   Grados de libertad del error (variación aleatoria) = [(número de factores A)(número de factores B)(número de réplicas menos 1)] = (3)(3)(2 – 1) = 9.

   Los correspondientes promedios de cuadrados:

$$PCFA = \frac{59.04}{2} = 29.52$$

$$PCFB = \frac{96.7}{2} = 48.35$$

$$PCAB = \frac{2\,072.89}{4} = 518.22$$

$$PCE = \frac{466}{9} = 51.78$$

5. Se obtienen los valores de F para cada caso, empíricos y teóricos. Como en los ejemplos anteriores, se utiliza un nivel de significación, alfa (α), de 0.05.

   Para el factor A, zonas:

$$F_e = \frac{PCFA}{PCE} = \frac{29.52}{51.78} = 0.57$$

   El valor crítico de la F, con dos grados de libertad para el numerador y 9 para el denominador es 4.26, ya que

$$P(F \geq 4.26 | gl_n = 2, gl_d = 9) = 0.05.$$

   Como el valor empírico de F, 0.57, es menor que este valor crítico de 4.26, no se rechaza la hipótesis nula referente a las zonas y se concluye que la media de las ventas en las 3 zonas son todas iguales.

   Para el factor B, vendedores:

$$F_e = \frac{PCFB}{PCE} = \frac{48.35}{51.78} = 0.93$$

   El valor crítico de la F, al igual que con el factor A, con 2 grados de libertad para el numerador y 9 para el denominador es 4.26, ya que

$$P(F \geq 4.26 | gl_n = 2, gl_d = 9) = 0.05$$

   Como el valor empírico de F, 0.93, es menor que este valor crítico de 4.26, no se rechaza la hipótesis nula referente a los vendedores y se concluye que la medias de las ventas de los 3 vendedores son todas iguales.

Para la interacción entre zonas y vendedores:

$$F_e = \frac{PCAB}{PCE} = \frac{518.22}{51.78} = 10.01$$

El valor crítico de la F, con 4 grados de libertad para el numerador y 9 para el denominador es 3.63, ya que

$$P(F \geq 3.63 | gl_n = 4, gl_d = 9) = 0.05.$$

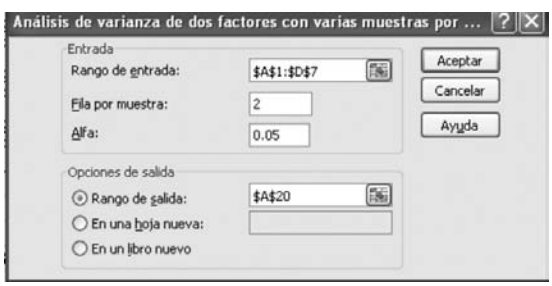
Como el valor empírico de F, 10.01, es mayor que este valor crítico de 3.63, se rechaza la hipótesis nula referente a la interacción y se concluye que sí existe interacción entre la zona y el vendedor.

Tal como en los casos anteriores, lo importante para el administrador es el uso que da a la información derivada del trabajo de análisis de varianza. El hecho de que los vendedores tengan diferentes desempeños en distintas zonas debe tener diversas causas, encontrarlas puede hacer más eficiente a la fuerza de ventas o puede ayudar a determinar un mejor perfil para los vendedores en distintas zonas.

12.10 Excel y ANOVA de dos factores con interacción

El complemento de Excel "Análisis de datos" también incluye una herramienta para este tipo de análisis de varianza, llamado *Análisis de varianza de dos factores con varias muestras por grupo*.

Si se colocan los datos en una hoja de Excel, a partir de la celda A1, incluyendo los títulos de renglón y de columna:



Profesión	Centro	Sur	Norte
Químico	506	528	513
	512	534	495
Lic en Admin.	529	496	508
	525	498	500
Ing. Mec.	500	512	528
	518	504	520

Después se elige este tipo de análisis desde la sección de "Análisis de datos", se llega al cuadro de diálogo que se muestra a la izquierda, en el cual puede verse que se eligen las celdas donde están los datos como "Rango de entrada", se le indica al programa que hay 2 datos para cada combinación de zona y vendedor en "Fila por muestra", se especifica el nivel de significación en "Alfa", 0.05, y se le pide colocar los resultados a partir de la celda A20, que son los siguientes:

Análisis de varianza de dos factores con varias muestras por grupo						
Resumen	**Centro**	**Sur**	**Norte**	**Total**		
Químico						
Cuenta	2	2	2	6		
Suma	1 018	1 062	1 008	3 088		
Promedio	509	531	504	514.666667		
Varianza	18	18	162	204.666667		
Licenciado en administración						
Cuenta	2	2	2	6		
Suma	1 054	994	1 008	3 056		
Promedio	527	497	504	509.333333		
Varianza	8	2	32	205.466667		
Ingeniero mecánico						
Cuenta	2	2	2	6		

Análisis de varianza de dos factores con varias muestras por grupo						
Resumen	**Centro**	**Sur**	**Norte**	**Total**		
Suma	1 018	1 016	1 048	3 082		
Promedio	509	508	524	513.666667		
Varianza	162	32	32	109.466667		
Total						
Cuenta	6	6	6			
Suma	3 090	3 072	3 064			
Promedio	515	512	510.666667			
Varianza	124	251.2	151.866667			
Análisis de varianza						
Origen de las variaciones	**Suma de cuadrados**	**Grados de libertad**	**Promedio de los cuadrados**	***F***	**Probabilidad**	**Valor crítico para *F***
Muestra	96.4444444	2	48.2222222	0.93133047	0.42892196	4.25649473
Columnas	59.1111111	2	29.5555556	0.57081545	0.58426005	4.25649473
Interacción	2 072.88889	4	518.222222	10.0085837	0.00227265	3.63308851
Dentro del grupo	466	9	51.7777778			
Total	2 694.44444	17				

Se invita al lector a verificar que se encuentran contenidos en esta tabla todos los datos necesarios para realizar las pruebas de las hipótesis planteadas, incluyendo el valor de probabilidad que permite realizarlas mediante lo que se llama aquí el *método de la P*.

ejercicios 12.10 ANOVA de dos factores con interacción

1. La gerencia de recursos humanos de una empresa que se dedica al diseño, construcción y venta de muebles para oficina recibió varias quejas de algunas empleadas, en relación con la existencia de una discriminación de género y que las empleadas recibían menos sueldo que sus compañeros varones por el mismo trabajo. Para determinar lo fundamentado de estas quejas, el gerente de recursos humanos recopiló información sobre una muestra elegida al azar de las compensaciones recibidas, en un mes dado, por los ocupantes de 3 puestos distintos y de diferente género, pero que se supone que tienen ingresos semejantes. Esta información se muestra a continuación. Las cifras son los ingresos mensuales de los ocupantes de cada uno de los puestos

a) ¿Podemos afirmar que efectivamente los puestos tienen ingresos semejantes y que ambos géneros tienen ingresos parecidos?

b) ¿Es posible afirmar que no existe interacción entre el género y el puesto?

Trabaje con alfa (α) de 0.05, plantee las hipótesis y realice los cálculos.

Escriba un reporte en el que le explique usted sus conclusiones al gerente de recursos humanos.

	Hombre	**Mujer**
Diseño de muebles	16 814	17 105
	16 438	15 418
	16 262	15 672
Ventas	17 910	16 571
	16 065	15 407
	16 854	15 573
Remodelación de interiores	17 093	17 391
	15 925	16 585
	16 657	15 734

2. Una empresa que vende ropa deportiva tiene tiendas en el centro de la ciudad, en los suburbios y en las plazas comerciales localizadas en diversos puntos; asimismo, la empresa tiene 2 sistemas de ventas: venta personalizada y autoservicio. Los datos sobre las ventas de varios meses elegidos al azar se encuentran en la siguiente tabla.

a) Trabaje con un alfa (α) de 0.01, plantee las hipótesis y realice los cálculos.
b) Responda:

- ¿Existirá diferencia en el volumen de ventas debida al lugar en el que se encuentra la tienda?
- ¿Existe diferencia entre los 2 sistemas de venta?
- ¿Se detecta interacción entre los sistemas y la localización de la tienda?

	Personalizada	Autoservicio
Centro	227 509	226 256
	220 362	227 657
	233 186	228 207
	226 876	209 436
Suburbios	211 387	217 615
	237 938	209 424
	246 110	208 563
	206 014	213 038
Plazas comerciales	223 66	216 007
	202 050	246 992
	245 545	201 877
	205 398	204 968

3. Una universidad evalúa a sus maestros cada semestre con un cuestionario que arroja una calificación de 0 a 100 puntos. Los maestros del turno matutino cubren los horarios de 7 a 9, de 9 a 11 y de 11 a 13 horas, los del turno vespertino laboran de 16 a 18, de 18 a 20 y de 20 a 22. Cuatro maestros cubrieron los 3 horarios del turno matutino y otros 4 los del turno vespertino durante varios semestres y la dirección desea saber:

a) Si los maestros, en promedio, tienen la misma calificación.
b) Si la evaluación de los docentes es la misma de horario en horario.
c) Si existe interacción entre la variable de horarios y la de los maestros.

Con base en la información que aparece en la siguiente tabla, haga un estudio de análisis de varianza que conteste estas preguntas y elabore un reporte para la dirección, en la que se expresen sus conclusiones, utilice un nivel de significación de 5 por ciento.

	Horario					
Maestro	7 a 9	9 a 11	11 a 13	16 a 18	18 a 20	20 a 22
Aguilar	79	96	86	93	82	88
	96	96	94	94	93	83
	95	99	79	96	94	90
Gutiérrez	78	98	93	75	94	88
	87	87	92	85	93	88
	87	90	90	78	92	80
Martínez	95	85	85	85	86	87
	82	85	87	90	93	96
	96	89	72	89	88	95
Rueda	99	90	89	85	82	92
	82	90	72	84	83	87
	95	86	72	92	100	94

4. Una empresa dedicada a diseñar y fabricar empaques especiales de cartón para productos del hogar tiene 3 máquinas distintas para imprimir la cara de los empaques que queda a la vista del público y las 3 pueden alimentarse con cartón nuevo blanqueado y no blanqueado o reciclado. La gerencia desea saber si las 3 máquinas producen el mismo porcentaje de piezas defectuosas, si el tipo de cartón tiene qué ver con el porcentaje de fallas y si existe interacción entre el tipo de cartón y la máquina que realiza la impresión. Para ello, se hacen varias corridas experimentales de prueba, los datos del porcentaje de fallas en estas corridas se encuentran en la tabla adjunta. ¿Cuál es su opinión sobre estos 3 aspectos utilizando un nivel de significación de 0.01?

	Tipo de cartón		
	Nuevo blanqueado	Nuevo no blanqueado	Reciclado
Impresora A	3.1	3.0	2.9
	3.3	2.9	2.7
	3.2	3.1	2.8
Impresora B	1.9	2.1	2.8
	2.3	2.0	2.6
	2.1	2.3	2.4
Impresora C	2.4	2.6	2.6
	2.7	2.5	2.2
	2.6	2.9	2.4

5. Para los siguientes datos de factores A y B y réplicas, pruebe las siguientes hipótesis:

a) Igualdad de medias de bloques.
b) Igualdad de medias de tratamientos.
b) Interacción, con un nivel de significación de 0.01.

	Factor A		
	A1	A2	A3
Factor B1	51	53	52
	63	62	66
Factor B2	61	60	62
	55	54	56

	Factor A		
	A1	A2	A3
Factor B3	57	56	52
	53	51	49

12.11 Resumen

En este capítulo se revisaron las principales técnicas del análisis de varianza, las cuales se usan para realizar pruebas de hipótesis sobre la igualdad de más de 2 medias poblacionales.

Las técnicas que se estudiaron fueron el ANOVA de un factor, el ANOVA de dos factores y el ANOVA de dos factores con interacción; se vio que el procedimiento de prueba consiste, básicamente, en calcular diferentes sumas de cuadrados y promedios de cuadrados para, con ellos, encontrar una F de Fischer empírica y, con base en el nivel de significación, una F crítica, de cuya comparación se desprende la aceptación o rechazo de la hipótesis nula y la consiguiente conclusión en términos del planteamiento del cual se parte.

Se vio también que, en el caso de este ANOVA, los mecanismos que contiene el complemento de Excel denominado "Análisis de datos" son de gran ayuda con los laboriosos cálculos que se requieren.

En los cuadros siguientes se resumen los principales elementos de este análisis de varianza, comenzando por el modelo que relaciona las diferentes sumas de cuadrados en cada caso.

Análisis de varianza de un factor.
$SCT = SCTr + SCE$

Suma de cuadrados	Grados de libertad
SCT	
$SCTr$	$gl_{Tr} = q - 1$
SCE	$gl_E = nq - q$

Análisis de varianza de dos factores.
$SCT = SCTr + SCB + SCE$

Suma de cuadrados	Grados de libertad
SCT	
$SCTr$	$gl_{Tr} = q - 1$
SCB	$gl_B = n - 1$
SCE	$gl_E = (q - 1)(n - 1)$

Análisis de varianza de dos factores con interacción.
$SCT = SCFA + SCFB + SCAB + SCE$

Suma de cuadrados	Grados de libertad
SCT	
$SCFA$	$gl_A = q - 1$
$SCFB$	$gl_B = n - 1$
$SCAB$	$gl_{AB} = (q - 1)(n - 1)$
SCE	$gl_E = qnr - 1$

Se revisó también, respecto al tema del análisis de varianza de un factor, el procedimiento que se sigue para hacer comparaciones entre pares de medias, cuando se da el caso de rechazar la hipótesis de igualdad de medias.

12.12 Fórmulas del capítulo

12.3 El diseño completamente aleatorio de un factor

El modelo de análisis de varianza de un factor:

$$SCT = SCTr + SCE$$

$$\sum_{i=1}^{n}\sum_{j=1}^{q}\left(X_{i,j} - \bar{X}_g\right)^2 = n\sum_{i=1}^{n}\left(\bar{X}_i - \bar{X}_g\right)^2 + \sum_{i=1}^{n}\sum_{j=1}^{q}\left(X_{i,j} - \bar{X}_i\right)^2 \quad (12.1)$$

12.4 Procedimiento para el ANOVA con el diseño completamente aleatorio de un factor

Media de tratamiento:

$$\bar{X}_{i,q} = \frac{\sum_{i=1}^{n} X_{i,q}}{n} \quad (12.2)$$

Media general o global:

$$\bar{X}_g = \frac{\sum_{i=1}^{n}\sum_{j=1}^{q} X_{i,q}}{nq} \quad (12.3)$$

La suma de cuadrados total:

$$SCT = \sum_{i=1}^{n}\sum_{j=1}^{q}\left(X_{i,j} - \bar{X}_g\right)^2 \quad (12.4)$$

La suma de cuadrados de tratamientos:

$$SCTr = n\sum_{i=1}^{n}\left(\bar{X}_{i,j} - \bar{X}_g\right)^2 \quad (12.5)$$

La suma de cuadrados del error:

$$SCE = \sum_{i=1}^{n} \sum_{j=1}^{q} \left(X_{i,j} - \bar{X}_i\right)^2 \quad (12.6)$$

Los promedios de cuadrados de tratamientos:

$$PCTr = \frac{SCTr}{gl_{Tr}} \quad (12.7)$$

Los promedios de cuadrados del error:

$$PCE = \frac{SCE}{gl_E} \quad (12.8)$$

El valor empírico del estadístico de prueba:

$$F_e = \frac{PCTr}{PCE} \quad (12.9)$$

12.13 Ejercicios adicionales

12.4 ANOVA con diseño completamente aleatorizado de un factor

1. Una armadora de automóviles tiene 5 concesionarias en la ciudad de Saltillo. Se supone que los técnicos de los talleres mecánicos de todas las concesionarias recibieron el mismo entrenamiento y, por lo mismo, los tiempos medios en los que realizan las tareas debe ser igual o muy similar; a continuación aparecen los tiempos (en minutos) que tardaron en hacer la afinación de una muestra de autos de 6 cilindros, del mismo modelo y año, los mecánicos de los distintos talleres. ¿Considera usted que los tiempos medios son efectivamente iguales? En caso de que no lo sean, ¿puede usted determinar cuál de los concesionarios es el más rápido y cuál el más lento? Utilice un alfa (α) de .01.

Concesionaria				
A	B	C	D	E
218	233	197	220	210
214	226	206	194	179
215	209	216	194	214
231	237	201	202	204
221	236	230	206	207
229	237	219	220	215
237	224	191	191	213

2. Los fabricantes de automóviles compactos de 4 marcas distintas promocionan sus autos como los más económicos en cuanto a consumo de combustible. Una asociación civil de defensa del consumidor supone que los rendimientos de los autos son, en promedio, iguales; para comprobar este supuesto, sujeta a una muestra de carros nuevos y en buenas condiciones mecánicas a un mismo recorrido mixto de ciudad y carretera, a las mismas horas del día y en el mismo día de la semana. Los resultados de rendimiento en km/L se muestran en la siguiente tabla. ¿Se puede sostener la afirmación de la asociación de que los autos tienen rendimientos aproximadamente iguales? En caso contrario, ¿se puede identificar cuál es la marca más eficiente? Utilice un alfa (α) de 0.05.

Marca			
A	B	C	D
16	18	22	20
19	19	21	17
18	17	20	16
16	18	19	20
17	18	21	16
20	21	20	20

3. Una universidad le hace examen de admisión a aspirantes de 3 sistemas educativos distintos: sistema escolarizado urbano, sistema escolarizado rural y sistema autodidacta por internet. La rectoría de la universidad está interesada en saber si existe alguna diferencia en cuanto a rendimiento en el examen de admisión entre los aspirantes a los 3 sistemas; para definirlo, obtiene una muestra aleatoria de estudiantes de los 3 sistemas que presentaron el último examen de admisión (la calificación máxima es de 200 puntos). ¿Considera usted que existe alguna diferencia? En caso afirmativo, ¿puede usted determinar cuál de los sistemas es el mejor o el peor? Utilice un alfa (α) de 0.01. Los rendimientos de la evaluación mencionada se dan a continuación.

Escolarizado urbano	Escolarizado rural	Autodidacta por internet
165	167	167
154	164	162
161	170	154
158	170	166
150	158	150
162	154	158
163	169	163
158	158	158
156	170	165
169	155	164

4. Una de las erogaciones que más preocupan a los gerentes de ventas es el de los gastos de viaje de los vendedores foráneos. Un gerente de ventas desea comparar el comportamiento de 3 vendedores distintos que atienden la misma zona. Para ello, obtuvo una muestra aleatoria de los gastos de viaje diarios de estos empleados (excluyendo transportación). Utilice un alfa (α) de 0.01 y responda lo siguiente:

a) ¿Es consistente el comportamiento de los 3 vendedores?
b) ¿Qué nuevas investigaciones le sugeriría usted al gerente?

Gómez	Gutiérrez	Hernández
650	490	570
620	685	520
510	710	780
640	540	670
550	620	520

5. Una planta de productos químicos tiene 4 plataformas de embarque idénticas para sus productos, que deben trabajar de manera uniforme. Existen también 4 supervisores que se rotan entre las plataformas todas las semanas. Los supervisores tienen la facultad de autorizar horas extra a su personal para realizar labores extraordinarias. Existe la preocupación en la gerencia de embarques de que alguno o algunos de los supervisores tengan más propensión a autorizar horas extra que otros; para investigarlo, se obtuvieron muestras aleatorias de las horas extra por semana autorizadas por cada uno de los supervisores. ¿Qué se puede concluir al respecto?, ¿los supervisores tienen un comportamiento uniforme o alguno autoriza más horas extra que los demás? Utilice un nivel de significación de 5 por ciento.

López	Mancilla	Ortiz	Oropeza
56	77	61	71
52	70	54	79
57	75	72	64
68	86	75	85

12.7 ANOVA de dos factores

6. Revise el problema 5 a partir de la idea de que pueden existir diferencias tanto en el comportamiento de los supervisores como en el hecho de que las plataformas de embarque pueden no ser idénticas. Realice un análisis de varianza de 2 factores e indique cuáles son sus conclusiones. ¿Los supervisores tienen, en promedio, el mismo comportamiento en relación con las horas extra?, ¿todas las plataformas son uniformes? Utilice un nivel de confianza de 95 por ciento.

Los datos son ahora:

	López	Mancilla	Ortiz	Oropeza
Plataforma 1	56	77	61	71
Plataforma 2	52	70	54	79
Plataforma 3	57	75	72	64
Plataforma 4	68	86	75	85

7. Una empresa analiza adquirir nueva máquina y existen 4 proveedores distintos para la misma. Se eligen aleatoriamente 5 de los operarios de la planta para trabajar con las diferentes máquinas durante una hora. El número de piezas producido por cada operador y por cada máquina se detalla a continuación. ¿Se puede concluir que todas las máquinas producen, en promedio, el mismo número de piezas?, ¿se puede concluir que todos los operadores tienen el mismo nivel de eficiencia? Use un nivel de confianza de 90 por ciento.

	Operario A	Operario B	Operario C	Operario D	Operario E
Máquina 1	82	78	82	79	79
Máquina 2	76	74	77	71	75
Máquina 3	84	81	82	79	85
Máquina 4	82	78	83	77	79

8. Una cadena de gasolineras tiene 3 establecimientos en la ciudad de Tuxtla Gutiérrez, uno en el centro de la ciudad, otro cercano al nuevo aeropuerto, y el último en una zona residencial. La gerencia desea saber si existe alguna diferencia entre los 3 locales y entre las ventas de los diferentes días de la semana. Para un nivel de significación de 0.05, indique si estas diferencias existen para alguna de las 2 variables: local y día de la semana. Las ventas se dan en litros diarios de gasolina.

	Local		
Día de la semana	Centro	Aeropuerto	Residencial
Lunes	7 404	7 557	7 598
Martes	7 515	7 693	7 512
Miércoles	7 670	7 480	7 602
Jueves	7 517	7 671	7 668
Viernes	7 199	7 155	7 140
Sábado	7 641	7 597	7 499
Domingo	7 131	7 129	7 131

9. Una empresa trabaja 3 turnos y ensambla 5 modelos de computadora que se numeran del 1 al 5. Las computadoras se prueban y, en caso de falla, deben volver a trabajarse hasta quedar en buen estado. La gerencia desea saber si existe alguna diferencia entre el porcentaje de fallas que se presentan en los diferentes modelos y entre los distintos turnos. ¿Qué puede usted informar a la gerencia al respecto con una significación de 0.01? Los datos de una muestra aleatoria de porcentajes de fallas por turno y modelo se muestran en la siguiente tabla.

	Turno		
Día de la semana	Matutino	Vespertino	Nocturno
Modelo 1	6.8	6.6	6
Modelo 2	6.7	5.6	6
Modelo 3	6.9	6.5	5.8
Modelo 4	6.7	5.9	5.5
Modelo 5	7.1	6.3	5.7

10. Una empresa tiene la concesión de la cafetería en varias empresas de distintos giros: computación, servicios de intermediación financiera y embotelladoras. Algunas de las empresas se sitúan en la zona residencial de la ciudad, otras

en el sector industrial y las últimas en la zona comercial de la misma. ¿Existe diferencia entre el consumo medio de estas cafeterías, según el giro de las empresas en donde se encuentran o de acuerdo con la zona de la ciudad? En la siguiente tabla se dan los consumos semanales, en pesos, elegidos al azar en diferentes empresas. Utilice un nivel de significación de 0.05.

	Giro de la empresa en donde se instaló la cafetería		
Tipo de zona	Computación	Intermediación	Embotelladoras
Zona residencial	26 570	28 470	28 508
Zona comercial	27 717	27 452	25 976
Zona industrial	25 983	28 649	26 038

12.9 ANOVA de dos factores con interacción

11. Para evaluar el efecto que tienen las proteínas, los carbohidratos y las grasas en el crecimiento de los ratones, se realizó un experimento con grupos de 3 ratones recién nacidos, de 3 razas distintas, a los que se les administró una dieta rica en cada uno de los nutrientes y se obtuvieron los siguientes resultados:

	Dieta rica en		
	Grasas (peso en gramos)	Proteínas (peso en gramos)	Carbohidratos (peso en gramos)
Grupo de ratones 1	25	35	12
	27	38	10
	28	36	9
Grupo de ratones 2	28	38	15
	30	41	13
	31	39	12
Grupo de ratones 3	18	25	8
	19	27	7
	20	25	6

Con un nivel de significación de 0.01, realice un análisis de varianza de dos factores con interacción.

12. Para evaluar el efecto que tienen ciertos tipos de carbohidratos y ciertos tipos de proteínas en la producción de leche de unas vacas, se realiza un diseño completamente al azar y se obtienen los siguientes resultados, en litros de leche por día:

	Tipo de carbohidratos		
Tipo de proteína	A	B	C
A	20	30	18
	25	35	23
B	15	11	13
	18	13	16
C	15	25	13
	20	30	18

Con un nivel de significación de 0.05, pruebe si hay diferencias entre las medias de los diferentes carbohidratos, de los distintos tipos de proteína y pruebe, además, si existe interacción entre estos 2 factores.

13. Para conocer el comportamiento del rendimiento de maíz como respuesta a la aplicación de diferentes niveles de riego y 3 dosis de nitrógeno aplicadas al suelo una semana antes de la siembra, se aplican estos factores a pares de parcelas y se obtienen los siguientes resultados, en kilogramos de maíz:

	Dosis de nitrógeno (kg)		
Niveles de riego	60	120	180
1	510	530	520
	550	540	560
2	630	620	660
	570	560	520
3	610	600	620
	530	510	490

a) Pruebe si existe diferencia entre los rendimientos según dosis de nitrógeno y según niveles de riego.
b) Evalúe también si existe interacción entre los 2 factores, con un nivel de significación de 0.05.

14. Se prueban los efectos de 3 diferentes mezclas de publicidad sobre las ventas de determinado producto junto con las posibles diferencias en las ventas en 4 distintos tipos de expendios en donde se vende ese producto. Los resultados que se obtuvieron fueron:

	Mezclas de publicidad		
Expendio	A	B	C
Tienda de abarrotes	60	40	40
	30	60	20
	10	30	60
Tienda de conveniencia	60	20	50
	80	70	30
	90	80	40
Supermercado chico	40	50	30
	20	20	20
	30	50	30
Supermercado grande	20	30	20
	30	90	40
	70	100	70

Pruebe, con un nivel de significación de 0.01:

a) Si existe diferencia entre los efectos de las 3 mezclas de publicidad sobre las ventas.
b) Si existe diferencia entre los 4 tipos de expendios sobre las ventas.
c) Si existe interacción entre esas 2 variables.

15. Un laboratorio farmacéutico analiza el efecto que tienen 3 tipos de almidón y 3 tipos de lubricante sobre la cohesión de cierto tipo de pastillas, para lo cual prepara corridas de 4 pastillas para cada combinación de almidón y lubricante, obteniendo los resultados que se muestran a continuación (en gramos por mm^3):

	Tipos de almidón		
Tipos de lubricante	**A**	**B**	**C**
X	30	41	32
	35	42	31
	37	43	45
	38	45	47
Y	36	40	48
	26	15	40
	25	12	41
	23	13	42
Z	21	18	43
	28	21	44
	29	19	39
	24	23	40

Pruebe si existen diferencias entre las resistencias para los diferentes tipos de almidón, para los distintos tipos de lubricante y pruebe si existe interacción entre estos 2 factores, con un nivel de significación de 0.05.

capítulo 13

Análisis de regresión y correlación lineal simple

Sumario

13.1 Ecuación y recta de regresión
13.2 Método de mínimos cuadrados
13.2.1 Derivación algebraica de las ecuaciones normales
13.2.2 Derivación de las ecuaciones normales mediante derivadas parciales
13.3 Determinación de la ecuación de regresión
13.3.1 Despeje simultáneo de *a* y *b* en las 2 ecuaciones normales
13.3.2 Resolución simultánea de las 2 ecuaciones normales
13.3.3 Resolución mediante sumas de cuadrados
13.3.4 Uso de Excel
13.4 Modelo de regresión y sus supuestos
13.5 Sumas de cuadrados en el análisis de regresión
13.6 Desviación estándar de regresión
13.7 Inferencias estadísticas sobre la pendiente β_1
13.7.1 Pruebas de hipótesis sobre la pendiente β_1
13.7.2 Estimación por intervalo de β_1
13.8 Uso de la ecuación de regresión para estimación y predicción
13.8.1 Estimación por intervalo de *y* para valores dados de *x*
13.8.2 Pronósticos de *y* para valores dados de *x*
13.9 Recapitulación del análisis de regresión lineal simple
13.10 Análisis de correlación
13.10.1 Coeficiente de correlación y Excel
13.10.2 Momento-producto de Pearson, otra manera de interpretar el coeficiente de correlación
13.10.3 Prueba de hipótesis sobre el coeficiente de correlación
13.10.4 Correlación serial o autocorrelación
13.11 Resumen
13.12 Fórmulas del capítulo
13.13 Ejercicios adicionales

Análisis de regresión simple. Estudia la relación entre 2 variables.

En el **análisis de regresión simple** se estudia la relación entre 2 variables. Por ejemplo, un caso muy sencillo en economía habla de que las unidades de un producto que se compran (la demanda) dependen del precio al que se ofrezcan (la oferta). La ley de la oferta y la demanda, planteada en términos muy sencillos, dice que conforme mayor es el precio de un artículo, menor es su demanda y viceversa: a menor precio mayor demanda. Éste es un ejemplo de la relación entre 2 variables que puede estudiarse mediante un análisis de regresión, en donde el precio sería una de las variables y la cantidad de unidades de demanda (o demanda a secas) la otra. En este caso, la variable de precio se denomina *variable independiente* representada mediante x y la cantidad de unidades de demanda se llama *variable dependiente* representada por y.

Análisis de regresión múltiple. Tiene solamente una variable dependiente y una independiente.

Cuando se tiene solamente una variable dependiente y una independiente se trata de un análisis de regresión simple; en contraposición, el **análisis de regresión múltiple** maneja una variable dependiente pero existen cuando menos 2 variables independientes.

En el ejemplo de la oferta y la demanda se tendría un caso de regresión múltiple si se analiza el efecto combinado que pudieran tener sobre la demanda las variables de precio, el producto interno bruto, el dinero en circulación y posiblemente otras más.

Otros ejemplos de pares de variables dependiente e independiente, que podrían relacionarse entre sí, son el número de vendedores y la cantidad de productos vendidos; los años de antigüedad en el empleo y el sueldo, entre otros.

Para evaluar estas relaciones, lo primero que se hace es trazar los valores correspondientes en un plano cartesiano (un plano de ejes vertical y horizontal), conocido como *diagrama de dispersión*.

Ejemplo 13.1

El gerente de un banco desea saber si puede considerarse que el ahorro de las familias (variable y) depende de sus ingresos (variable x). En la tabla 13.1 se muestran los resultados obtenidos para una muestra de 10 familias.

Tabla 13.1 Ingreso y ahorro mensual de 10 familias (en miles de pesos).

Familia	Ingresos (x)	Ahorro (y)
1	11	0.5
2	14	1.1
3	12	0.9
4	9	0.6
5	13	1.2
6	13	0.9
7	15	1.5
8	17	1.3
9	15	1.1
10	13	0.7

La representación de estos datos en un diagrama de dispersión se ilustra en la figura 13.1.

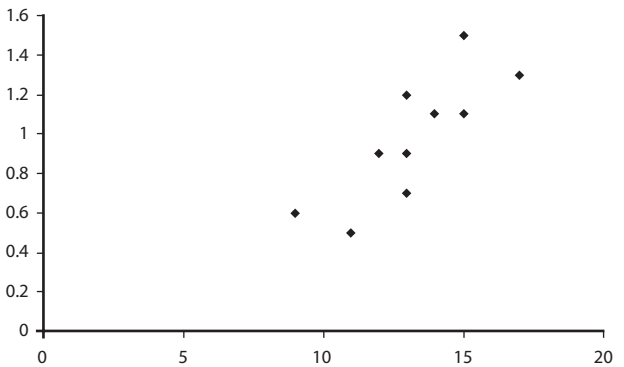


Figura 13.1 Diagrama de dispersión para los datos del ejemplo 13.1.

La figura 13.1 muestra un ejemplo de regresión donde la relación entre las variables parece ser rectilínea (lineal) y directa (al aumentar el valor de la variable independiente, aumenta el valor de la dependiente). En la figura 13.2 se presentan ilustraciones de los otros posibles tipos de relación entre 2 variables.

La figura 13.2*a*) es otro ejemplo de relación lineal positiva, mientras que la figura 13.2*b*) es un ejemplo de la relación lineal inversa (negativa), ya que puede apreciarse que al aumentar la variable independiente, disminuye el valor de la variable dependiente. En la figura 13.2*c*) se muestra el caso de una relación que no es lineal (se aprecia fácilmente que la relación es curvilínea). Finalmente, en la figura 13.2*d*) se ilustra el caso de un diagrama de dispersión que muestra la nula relación entre las 2 variables utilizadas para elaborar el diagrama.

En cualquier análisis de regresión simple conviene, como primer paso, trazar el diagrama de dispersión para darse una idea de si existe relación entre las variables y qué forma tiene ésta.

El análisis de regresión lineal simple tiene como objeto analizar relaciones como las ilustradas en los incisos *a*) y *b*) de la figura 13.2. Tiene la finalidad de determinar la ecuación de la recta que mejor describe la relación entre las 2 variables. Por su parte, el análisis de correlación se ocupa de evaluar el sentido y la intensidad de la relación, temas que se abordarán en la sección 13.11.

En el siguiente punto se muestra la forma en la que puede ajustarse a mano alzada una recta a una nube de puntos o diagrama de dispersión para encontrar la ecuación lineal que la representa.

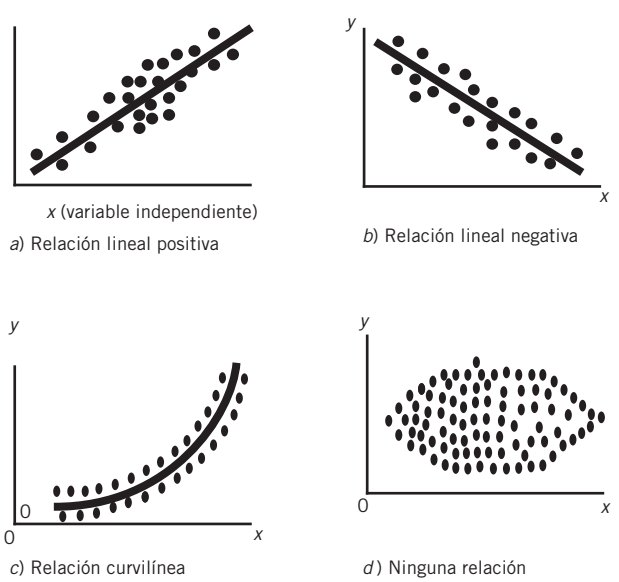


Figura 13.2 Ejemplos de diagramas de dispersión para datos con 2 variables.

13.1 Ecuación y recta de regresión

La forma más sencilla de ajustar una recta a los datos graficados en un diagrama de dispersión sería colocar una regla sobre la gráfica e intentar dibujar una línea que pase entre los puntos, de manera que éstos queden lo más cerca posible de la recta. Por supuesto, diferentes personas o la misma persona en diferentes

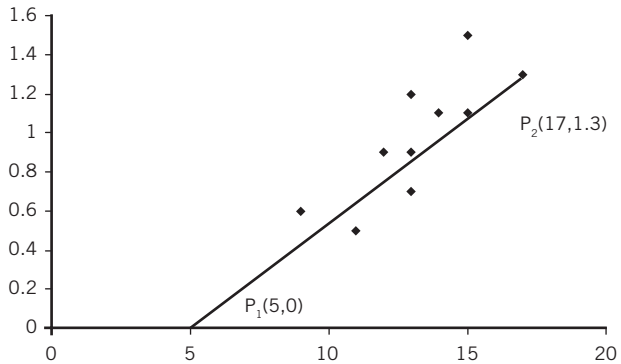


Figura 13.3 Ajuste a mano alzada de una recta al diagrama de dispersión de los datos de ingreso y ahorro.

momentos podrían dibujar rectas distintas mediante este procedimiento; sin embargo este ensayo permite visualizar qué se hace con el método matemático de mínimos cuadrados, el cual se utilizará y revisará con frecuencia más adelante.

Por lo pronto, en la figura 13.3 con este método manual se ha ajustado una recta a los datos de la figura 13.1. Trazada la recta, la manera de encontrar la ecuación correspondiente es identificando las coordenadas de 2 puntos y sustituyendo en la familiar fórmula:

$$\frac{y_2 - y_1}{x_2 - x_1} = \frac{y - y_1}{x - x_1} \qquad \textbf{(13.1)}$$

En la figura 13.3 se identifican los 2 puntos:

$$P_1(5,0) \quad \text{y} \quad P_2\,(17,1.3)$$

Con estas coordenadas, sustituyendo en la fórmula anterior:

$$\frac{1.3 - 0}{17 - 5} = \frac{y - 0}{x - 5}$$

Simplificando:

$$\frac{1.3}{12} = \frac{y}{x - 5}$$

$$1.08333 = \frac{y}{x - 5}$$

$$y = 1.0833\,(x - 5) = 1.0833x - 5.42$$

En resumidas cuentas:

$$y = 1.0833x - 5.42$$

Entonces, la ecuación anterior es la recta de regresión ajustada manualmente a los datos y puede verse que tiene la forma de la ecuación de pendiente y ordenada al origen de una función lineal:

$$y = mx + b$$

En donde m es la pendiente o inclinación de la recta, y b es la ordenada al origen; en otras palabras, la altura en donde la recta cruza el eje vertical.

En seguida se presenta otra manera de representar estas ecuaciones lineales, equivalente a la anterior y más conveniente para el formato en el que suelen presentarse las ecuaciones en el método de mínimos cuadrados:

$$y = a + bx \qquad \textbf{(13.2)}$$

en donde a es la ordenada al origen y b es la pendiente.

Una de las principales aplicaciones prácticas de estas rectas de regresión consiste en utilizarlas para estimar valores de la variable dependiente, con base en determinados valores de la variable dependiente, como en el ejemplo siguiente.

EJEMPLO 13.2

Utilizando la ecuación de la recta que se encontró antes, estime la cantidad de ahorro para una familia que tiene ingresos de $10 000 mensuales.

Solución: Sustituyendo en la ecuación de regresión:

$$y = 1.0833x - 5.42 = 1.0833(10) - 5.42 = 10.833 - 5.42 = 5.413$$

Así se estimaría que una familia con ingresos mensuales de $10 000 ahorraría $5 413 al mes.

ejercicios 13.1 Ecuación y recta de regresión

1. Se tiene una muestra de 8 estudiantes con 2 datos para cada uno: el número de horas que han dedicado al estudio de una asignatura y la calificación obtenida en el examen correspondiente: (38,7), (36,7), (42,9), (25,3), (27,4), (32,6), (48,9) y (40,5).

a) Dibuje el diagrama de dispersión de estos puntos.
b) Trace a mano alzada la recta que mejor se ajuste a los puntos.
c) Encuentre la ecuación de esa recta.
d) Estime la calificación que podría obtener un alumno que estudia 30 horas.

2. A continuación se muestran los datos de ventas anuales de una empresa manufacturera:

Año	Ventas (millones de pesos)
2002	1.5
2003	1.8
2004	1.6
2005	2.1
2006	2.6
2007	2.7
2008	2.5
2009	3.2
2010	3.1
2011	3.4

a) Dibuje el diagrama de dispersión de estos puntos.
b) Trace a mano alzada la recta que mejor se ajuste a los puntos.
c) Encuentre la ecuación de esa recta.
d) Estime las ventas de la empresa para 2012.

3. Para la ecuación: $y = 4x - 3$

a) Trace la recta correspondiente en un par de ejes coordenados.
b) Identifique la pendiente y la ordenada al origen.

4. Para la ecuación: $2x + 3y = 10$

a) Trace la recta correspondiente en un par de ejes coordenados.
b) Identifique la pendiente y la ordenada al origen.

5. Encuentre la ecuación de la siguiente recta a partir de las coordenadas de los 2 puntos proporcionados: $P_1(-3,-2)$ y $P_2(4,5)$.

6. Encuentre la ecuación de la siguiente recta a partir de las coordenadas de los 2 puntos proporcionados: $P_1(3,12)$ y $P_2(8,2)$.

7. ¿Qué diferencia entre las ecuaciones 3 y 4, y entre las 5 y 6 hace que las rectas tengan inclinaciones diferentes?

8. Dibuje la recta que cruza el eje vertical en el valor –3, y que tiene como pendiente 5. Encuentre también la ecuación correspondiente.

9. ¿Cuál es la principal desventaja de ajustar manualmente una recta a un diagrama de dispersión?

13.2 Método de mínimos cuadrados

El procedimiento manual que se explicó antes para ajustar una recta de regresión a un conjunto de datos trazados en un diagrama de dispersión, tiene la desventaja de ser bastante inexacto. El método que comúnmente se utiliza para realizar este ajuste es el llamado **método de mínimos cuadrados**.

Método de mínimos cuadrados. Reduce al mínimo el cuadrado de las distancias verticales entre cada uno de los puntos y la recta ajustada.

Se llama así porque reduce al mínimo el cuadrado de las distancias verticales entre cada uno de los puntos y la recta ajustada. La idea se aprecia visualmente en la figura 13.4.

Esta figura es la misma de la sección anterior del ejemplo de la relación entre ingresos y ahorro. Ahora se han marcado líneas verticales entre algunos de los puntos y la recta que representa su relación.

Cuando de manera manual se ajusta la recta al diagrama de dispersión, normalmente se intenta ubicarla lo más centrada posible en la nube de puntos. Esta forma intuitiva de colocar la recta pretende minimizar la distancia vertical entre cada uno de los puntos y la recta. Es precisamente esto lo que se hace cuando se utiliza el método de mínimos cuadrados, sin embargo con este método, por ser un procedimiento numérico, se garantiza que la suma de los cuadrados de esas desviaciones es mínima.

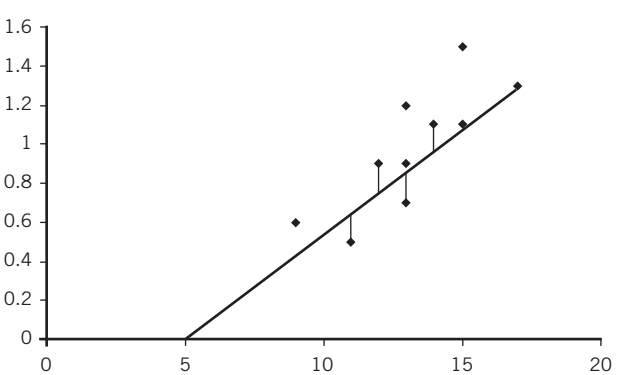


Figura 13.4 Distancias entre la ordenada de cada punto y la ordenada de la recta de regresión: $y_i - y_c$.

Ecuaciones normales. Conjunto de 2 ecuaciones que resueltas simultáneamente producen el valor de la pendiente y el de la ordenada al origen que son los parámetros de la ecuación lineal que arroja los mínimos cuadrados.

Este método se basa en lo que se conoce como las **ecuaciones normales**, un conjunto de 2 ecuaciones que resueltas simultáneamente (como se explicará más adelante) producen el valor de la pendiente y el de la ordenada al origen que son los parámetros de la ecuación lineal que arroja los mínimos cuadrados. En las secciones siguientes se mostrarán los procedimientos algebraicos y de cálculo diferencial mediante los cuales se obtienen las ecuaciones normales.

13.2.1 Derivación algebraica de las ecuaciones normales

Las 2 ecuaciones normales pueden obtenerse algebraicamente de la siguiente manera:

Si $y_1, y_2, \ldots y_n$ representan las observaciones de la variable y, $x_1, x_2, \ldots x_n$ representan las observaciones de la variable x.

Entonces se tiene que:

$$y_1 = a + bx_1,$$

$$y_2 = a + bx_2, \ldots$$

La suma de estas últimas es:

$$y_1 = a + bx_1$$

$$+$$

$$y_2 = a + bx_2$$

$$+$$

$$\ldots\ldots = \ldots + \ldots\ldots$$

$$+$$

$$y_n = a + bx_n$$

Resumidamente igual a:

$$\sum y = na + b\sum x$$

Si ahora se multiplica cada una de las ecuaciones de la forma $y = a + bx$ por el coeficiente de la primera incógnita de la ecuación (1 que es el coeficiente de la primera incógnita, a) no las altera. Así que la suma de esas ecuaciones resultantes (que no cambiaron) es igual a la suma anotada antes:

$$\sum y = na + b\sum x \tag{13.3}$$

Conocida como la *ecuación normal I*.

Si ahora se multiplica cada una de las ecuaciones de la forma $y = a + bx$ por el coeficiente de la segunda incógnita de la ecuación (x que es el coeficiente de la segunda incógnita, b) se obtiene:

$$x_1y_1 = ax_1 + bx_1^2$$

$$+$$

$$x_2y_2 = ax_2 + bx_2^2$$

$$+$$

$$\ldots\ldots = \ldots + \ldots\ldots$$

$$+$$

$$x_ny_n = ax_n + bx_n^2$$

Sumadas producen:

$$\sum xy = a\sum x + b\sum x^2 \tag{13.4}$$

Lo que se conoce como la *ecuación normal II*.

En resumen, el conjunto de 2 ecuaciones normales que resueltas simultáneamente permiten construir la ecuación lineal de mínimos cuadrados son:

$$\sum y = na + b\sum x$$
$$\sum xy = a\sum x + b\sum x^2$$

En la sección 13.3 se ilustrará su uso.

A continuación se revisará el procedimiento de cálculo diferencial mediante el cual se derivan estas mismas ecuaciones; además permite probar que, efectivamente, la suma de esas distancias verticales entre cada punto y la recta es mínima.

13.2.2 Derivación de las ecuaciones normales mediante derivadas parciales

Propondría cambiar la palabra derivación por obtención. En seguida se detalla la forma como se obtienen las ecuaciones normales pero ahora utilizando el cálculo diferencial, el cual permite probar que es mínima la suma de esas distancias verticales entre cada punto y la recta.

Se utiliza la siguiente simbología:

n = número de pares de valores observados de x y y.
y = valor observado de la variable dependiente.
x = valor observado de la variable independiente.
$\hat{y}$ = valor calculado de la variable dependiente de acuerdo con la ecuación de regresión.
$\hat{y} = a + bx$ es la ecuación de la recta de regresión que se busca, la que minimiza los cuadrados de las desviaciones.
$e = y - \hat{y}$ o sea la diferencia entre la ordenada (y) del punto observado y el valor de $\hat{y}$ correspondiente a la recta de regresión, es decir, las distancias verticales cuya suma se desea minimizar. La e es porque se le denomina *error*.

Entonces, mediante el método de mínimos cuadrados se minimiza:

$$\sum e^2 = \sum \left(y - \hat{y}\right)^2$$

A esta expresión suele llamársele, apropiadamente, la SCE o suma de los cuadrados de los errores.

Sustituyendo $\hat{y} = a + bx$ en la fórmula de la SCE se obtiene:

$$SCE = \sum e^2 = \sum \left(y - \hat{y}\right)^2 = \sum \left(y - a - bx\right)^2$$

Ya que se tienen 2 incógnitas, a y b, para minimizar esta función se obtienen las derivadas parciales con respecto a cada una de ellas:

I. La derivada parcial de la SCE con respecto a a:

$$\frac{\partial(SCE)}{\partial a} = \sum 2(y - a - bx)(-1) = 2\sum (-y + a + bx)$$

El punto crítico (mínimo) de esta función se encuentra en donde la derivada parcial es igual a cero:

$$\frac{\partial(SCE)}{\partial a} = 2\sum(-y + a + bx) = 0$$

Simplificando:

$$\sum(-y + a + bx) = 0$$
$$-\sum y + na + b\sum x = 0$$
$$\sum y = na + b\sum x \qquad \textbf{(I)}$$

Se trata de la misma ecuación normal I que previamente se encontró algebraicamente.

II. La derivada parcial de la SCE con respecto a b:

$$\frac{\partial(SCE)}{\partial b} = 2\sum(y - a - bx)(-x) = 2\sum(-xy + ax + bx^2)$$

Cuando esta derivada parcial es igual a 0, que es también un punto mínimo, se tiene:

$$2\sum(-xy + ax + bx^2) = 0$$
$$\sum(-xy + ax + bx^2) = 0$$
$$\sum(-xy) + a\sum x + b\sum x^2 = 0$$
$$\sum xy = a\sum x + b\sum x^2$$

Es la misma ecuación normal 2 que previamente se encontró algebraicamente.

Como se mencionó, resolviendo en forma simultánea este par de ecuaciones normales se obtiene la ecuación de regresión:

$$\sum y = na + b\sum x \qquad \textbf{(I)}$$
$$\sum xy = a\sum x + b\sum x^2 \qquad \textbf{(II)}$$

13.3 Determinación de la ecuación de regresión

En esta parte se ilustran 4 formas para calcular la ecuación de regresión lineal:

1. Despejando simultáneamente las incógnitas *a* y *b* en las 2 ecuaciones normales.
2. Resolviendo simultáneamente las 2 ecuaciones normales.
3. Utilizando promedios y sumas de cuadrados.
4. Utilizando Excel.

Estos procedimientos conducen a encontrar los parámetros de la ecuación de regresión y se incluyen aquí para una mejor comprensión del tema; aunque en la práctica conviene utilizar el más sencillo de ellos que puede ser a través de promedios y sumas de cuadrados, o utilizando el paquete Excel de Microsoft.

13.3.1 Despeje simultáneo de *a* y *b* en las 2 ecuaciones normales

Un primer método para construir la ecuación de regresión consiste en despejar las 2 incógnitas, *a* y *b*, la ordenada al origen y la pendiente de la ecuación de regresión. A partir de las 2 ecuaciones normales:

$$\sum y = na + b\sum x \qquad \textbf{(I)}$$
$$\sum xy = a\sum x + b\sum x^2 \qquad \textbf{(II)}$$

Dividiendo la ecuación I entre *n*:

$$\frac{\sum y}{n} = a + b\frac{\sum x}{n}$$
$$\hat{y} = a + b\bar{X}$$
$$a = \hat{y} - b\bar{X}$$

Sustituyendo este resultado en la ecuación II:

$$\sum xy = (\tilde{y} - b\bar{x})\sum x + b\sum x^2$$
$$\sum xy = \bar{y}\sum x - b\bar{x}\sum x + b\sum x^2$$
$$\sum xy = \bar{y}\sum x - b\left(\bar{x}\sum x - \sum x^2\right)$$
$$\sum xy = \frac{\sum y}{n}\sum x - b\left(\frac{\sum x}{n}\sum x - \sum x^2\right)$$
$$\sum xy = \frac{\sum y\sum x}{n} - b\left(\frac{\left(\sum x\right)^2}{n} - \sum x^2\right)$$
$$b\left(\frac{\left(\sum x\right)^2}{n} - \sum x^2\right) = \frac{\sum y\sum x}{n} - \sum xy$$

Por lo que:

$$b = \frac{\dfrac{\sum y \sum x}{n} - \sum xy}{\dfrac{\left(\sum x\right)^2}{n} - \sum x^2}$$

Al multiplicar el numerador y el denominador del segundo término por –1 se obtiene la forma de esta expresión frecuentemente utilizada:

$$b = \frac{\sum xy - \dfrac{\sum x \sum y}{n}}{\sum x^2 - \dfrac{\left(\sum x\right)^2}{n}}$$

Resumiendo el análisis anterior, para encontrar los parámetros a y b de la recta de regresión se utilizan las siguientes expresiones:

$$a = \bar{y} - b\bar{x}, \text{ y} \tag{13.5}$$

$$b = \frac{\sum xy - \dfrac{\sum x \sum y}{n}}{\sum x^2 - \dfrac{\left(\sum x\right)^2}{n}} \tag{13.6}$$

Ejemplo 13.3

Calcule la ecuación de regresión para los datos de ingreso y ahorro del ejemplo 13.1.

Solución: En la tabla 13.2 se reproducen los datos así como los cálculos necesarios para determinar la ecuación.

Tabla 13.2 Datos de ingreso y ahorro; cálculos para determinar la ecuación de regresión de mínimos cuadrados.

	Ingreso x	Ahorro y	xy	x^2
1	11	0.5	5.5	121
2	14	1.1	15.4	196
3	12	0.9	10.8	144
4	9	0.6	5.4	81
5	13	1.2	15.6	169
6	13	0.9	11.7	169
7	15	1.5	22.5	225
8	17	1.3	22.1	289
9	15	1.1	16.5	225
10	13	0.7	9.1	169
Totales	132	9.8	134.6	1 788

Sustituyendo en las fórmulas que se despejaron antes para a y b:

$$b = \frac{\sum xy - \dfrac{\sum x \sum y}{n}}{\sum x^2 - \dfrac{\left(\sum x\right)^2}{n}} = \frac{134.6 - \dfrac{132(9.8)}{10}}{1\,788 - \dfrac{132^2}{10}} = \frac{134.6 - 1\,29.36}{1\,788 - 1\,742.4}$$

$$= \frac{5.24}{45.6} = 0.1149$$

La sustitución en a:

$$a = \bar{y} - b\bar{x} = \frac{9.8}{10} - 0.1149\frac{132}{10} = 0.98 - 1.5167 = -0.5367$$

De donde la ecuación de regresión lineal es:

$$\hat{y} = -0.5367 + 0.1149x$$

13.3.2 Resolución simultánea de las 2 ecuaciones normales

El procedimiento anterior, que consiste en sustituir los valores de sumatorias y promedios en las ecuaciones para a y para b despejadas de las 2 ecuaciones normales, es una forma abreviada del procedimiento más general que implica sustituir los valores de esas sumatorias en las ecuaciones normales y luego resolverlas simultáneamente. Se ilustra en seguida este último procedimiento.

Sustituyendo en las ecuaciones normales:

$$\sum y = na + b\sum x$$

$$\sum xy = a\sum x + b\sum x^2$$

$$\text{I. } \sum y = na + b\sum x$$
$$9.8 = 10a + 132b$$
$$\text{II. } \sum xy = a\sum x + b\sum x^2$$
$$134.6 = 132a + 1\,788b$$

Resolviendo estas ecuaciones en forma simultánea:

$$\text{I. } 9.8 = 10a + 132b$$
$$\text{II. } 134.6 = 132a + 1\,788b$$

Despejando a en I:

$$10a = -132b + 9.8$$
$$a = \frac{-132b + 9.8}{10}$$
$$a = -13.2b + 0.98$$

Sustituyendo este resultado en II:

$$134.6 = 132\,(-13.2b + 0.98) + 1\,788b$$
$$134.6 = -1\,742.4b + 129.36 + 1\,788b$$
$$134.6 = 45.6b + 129.36$$
$$45.6b = 134.6 - 129.36 = 5.24$$
$$b = \frac{5.24}{45.6}$$
$$b = 0.1149$$

De vuelta a la ecuación I se sustituye este valor de b:

$$a = -13.2b + 0.98$$
$$a = -13.2(0.1149) + 0.98$$
$$= -1.5167 + 0.98$$
$$a = -0.5367$$

Por lo que la ecuación de regresión de mínimos cuadrados es:

$$\hat{y} = -0.5367 + 0.1149x$$

Que es la misma ecuación encontrada antes.

13.3.3 Resolución mediante sumas de cuadrados

Otra manera de encontrar la pendiente y la ordenada al origen de la ecuación de regresión de mínimos cuadrados es a través de las siguientes fórmulas:

La misma ecuación (13.6) que se derivó antes:

$$a = \bar{y} - b\bar{x} \tag{13.5}$$

Y:

$$b = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2} = \frac{SC_{xy}}{SC_{xx}} \tag{13.7}$$

El ejemplo anterior ahora se resuelve utilizando este procedimiento. En la tabla 13.3 se muestran los datos y las operaciones.

$$SC_{xy} = 5.24$$
$$SC_{xx} = 45.6$$

De donde:

$$b = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2} = \frac{SC_{xy}}{SC_{xx}} = \frac{5.24}{45.6} = 0.1149$$

Tabla 13.3 Datos y operaciones para resolver el ejercicio de ingreso y ahorro utilizando sumas de cuadrados.

	Ingreso	Ahorro		
	x	y	$(x-\bar{x})^2$	$(x-\bar{x})(y-\bar{y})$
1	11	0.5	4.84	1.056
2	14	1.1	0.64	0.096
3	12	0.9	1.44	0.096
4	9	0.6	17.64	1.596
5	13	1.2	0.04	–0.044
6	13	0.9	0.04	0.016
7	15	1.5	3.24	0.936
8	17	1.3	14.44	1.216
9	15	1.1	3.24	0.216
10	13	0.7	0.04	0.056
Sumas	132	9.8	45.6	5.24

Y:

$$a = \bar{y} - b\bar{x} = \frac{9.8}{10} = 0.1149\frac{132}{10} = 0.98 - 0.1149(13.2) = -0.5367$$

Se trata de los mismos valores obtenidos por los 2 métodos ilustrados previamente, y que producen la misma ecuación de regresión:

$$\hat{y} = -0.5367 + 0.1149x$$

Por supuesto, cada quien deberá utilizar el método que le parezca más conveniente; se sugiere emplear el más sencillo que parece ser este último. Anteriormente se utilizaban los 2 primeros métodos y además se aconsejaban los *métodos simplificados* de cálculo pero se han vuelto innecesarios dada la amplia disponibilidad de calculadoras electrónicas y computadoras que facilitan cualquier tipo de cálculo. Estas novedades se ilustran en la sección siguiente con el uso de Excel.

13.3.4 Uso de Excel

En este libro se ha utilizado el complemento de "Análisis de datos" de Excel, al que puede accederse desde la pestaña "Datos" ubicada en el extremo izquierdo de la cinta de opciones. Al dar clic aquí, aparece el menú de técnicas de análisis de datos que incluye la de "Regresión". Cuando se hace clic en esta opción aparece la pantalla reproducida en la figura 13.5.

Como puede verse, deberán anotarse los rangos de la hoja de Excel en las casillas de y y de x. Se aprecian otras opciones cuyo análisis se explicará más adelante porque ayudan a resolver aplicaciones de regresión lineal múltiple, el tema del siguiente capítulo.

Así que de momento basta con anotar en una hoja de Excel los datos de ingreso y ahorro con sus correspondientes encabezados; también debe activarse la casilla "Rótulos" en este cuadro de diálogo de Excel para indicar al programa que se incluyeron.

De esta manera sólo resta activar la casilla "Rango de salida" en la sección "Opciones de salida" y marcar alguna celda vacía ubicada hacia abajo y a la derecha para que

Figura 13.5 Pantalla para el uso de "Regresión" en Excel.

a partir de ella aparezcan los datos. Finalmente, se pulsa la tecla aceptar y se obtienen los resultados que se muestran en la tabla 13.4.

Tabla 13.4 Resultados de Excel para el ejemplo de ingreso y ahorro.

Estadísticas de la regresión								
Coeficiente de correlación múltiple				0.81077646				
Coeficiente de determinación R^2				0.65735846				
R^2 ajustado				0.61452827				
Error típico				0.19807185				
Observaciones				10				
Análisis de varianza								
	Grados de libertad	Suma de cuadrados	Promedio de los cuadrados	*F*	Valor crítico de *F*			
Regresión	1	0.60214035	0.60214035	15.3480157	0.00443301			
Residuos	8	0.31385965	0.03923246					
Total	9	0.916						
	Coeficientes	Error típico	Estadístico *t*	Probabilidad	Inferior 95%	Superior 95%	Inferior 95.0%	Superior 95.0%
Intercepción	–0.53684211	0.3922149	–.36874481	0.2082762	–1.44129128	0.36760707	–1.44129128	0.36760707
Ingreso	0.11491228	0.02933191	3.91765436	0.00443301	0.04727278	0.18255178	0.04727278	0.18255178

Esta tabla de resultados contiene mucha más información que aquella que se ha revisado hasta aquí. De momento sólo se contemplan los conceptos que se han aprendido, ya que en el capítulo siguiente se completa la explicación total de la tabla.

De la primera sección, "Estadísticas de regresión", por ahora sólo se considerarán las "Observaciones" en donde se anota que son 10 los pares de datos x con los que se realiza el ejemplo.

En la sección siguiente titulada "Análisis de varianza" se anotan, entre otros, los datos de sumas y promedios de cuadrados. Posteriormente se explicará con mayor detalle por qué se le llama análisis de varianza; de entrada puede apreciarse la similitud con esa técnica estadística revisada en el capítulo anterior con respecto a sumas y promedios de cuadrados.

Intercepción. Es la misma ordenada al origen, es decir, el punto en el que la recta cruza (intercepta) al eje vertical.
Ingreso. Marca la pendiente, lo cual lleva a la misma ecuación de regresión pero más precisa, con mayor número de decimales.

Finalmente, en los últimos renglones de la tabla 13.4 en la columna "Coeficientes" pueden apreciarse los valores –0.53684211 y 0.11491228, catalogados como intercepción e ingreso, respectivamente. **Intercepción** es la misma ordenada al origen, es decir, el punto en el que la recta cruza (intercepta) al eje vertical, sólo que con mayor precisión que la obtenida antes. En tanto que **ingreso** marca la pendiente, lo cual lleva a la misma ecuación de regresión pero más precisa, con mayor número de decimales:

$$\hat{y} = -0.53684211 + 0.11491228x$$

ejercicios 13.3 Determinación de la ecuación de regresión

1. De una muestra de 15 embarques de láminas de aluminio se registró la distancia en km al lugar de entrega así como el tiempo en horas para cada carga.

Embarque	Distancia en km	Tiempo de entrega en horas
1	400	7
2	800	12

Embarque	Distancia en km	Tiempo de entrega en horas
3	120	2
4	340	5
5	520	8
6	300	4
7	100	1.4
8	85	1
9	589	9.6
10	1 115	13
11	265	3
12	670	11
13	1 215	15
14	550	9
15	215	2.6

a) Elabore el diagrama de dispersión.
b) Determine ecuación de regresión con cualquiera de los 3 métodos presentados.
c) Grafique la recta de regresión en el diagrama de dispersión.
d) Verifique con Excel que los cálculos realizados sean correctos.

2. En un ingenio azucarero se aplicaron diferentes cantidades de fertilizantes en ciertos sectores de las 50 hectáreas de cultivo. Se tomó una muestra de 8 sectores; los resultados con respecto a la producción de caña se muestran en la siguiente tabla.

Sector	Fertilizante (kg)	Producción de caña (toneladas)
1	2	4
2	5	9
3	7	11
4	8	13
5	2.5	5
6	6	11
7	11	15
8	15	17

a) Elabore el diagrama de dispersión.
b) Determine ecuación de regresión con cualquiera de los 3 métodos presentados.
c) Grafique la recta de regresión en el diagrama de dispersión.
d) Verifique con Excel que los cálculos realizados sean correctos.

3. En una planta ensambladora de aparatos electrodomésticos se tomó una muestra de 20 trabajadores. En la siguiente tabla se muestran las semanas de experiencia con que cuenta cada empleado y el número de artículos que fueron rechazados, en una semana, por algún defecto en el armado

Trabajador	Semanas de experiencia	Productos rechazados
1	1	20
2	5	14
3	4	14
4	3	15
5	6	12
6	11	6
7	10	8
8	8	9
9	3	15
10	6	10
11	12	4
12	9	8
13	7	11
14	15	3
15	1	18
16	2	16
17	14	4
18	2	15
19	4	13
20	5	12

a) Elabore el diagrama de dispersión.
b) Determine ecuación de regresión con cualquiera de los 3 métodos presentados.
c) Grafique la recta de regresión en el diagrama de dispersión.
d) Verifique con Excel que los cálculos realizados sean correctos.

4. Después de un examen de matemáticas se preguntó a 10 estudiantes el número de horas que habían estudiado y la calificación que obtuvieron. Los resultados se presentan a continuación.

Estudiante	Horas de estudio	Calificación
1	10	9.8
2	5	9.1
3	1	6.2
4	2.5	7.3
5	8	9.6
6	4	9.2
7	3	8.8
8	7	9.3
9	6.5	9.5
10	4	9

a) Elabore el diagrama de dispersión.
b) Determine ecuación de regresión con cualquiera de los 3 métodos presentados.
c) Grafique la recta de regresión en el diagrama de dispersión.
d) Verifique con Excel que los cálculos realizados sean correctos.

5. De una muestra de 11 vendedores que trabajan para una empresa, se registraron los años de experiencia que tienen y el volumen de ventas que cada uno obtuvo durante el primer semestre del año.

Vendedor	Experiencia	Volumen de ventas (miles)
1	8	27
2	5	25
3	1	18
4	3	21
5	2	19
6	6	25
7	9	29
8	4	21
9	3	20
10	7	26
11	10	32

a) Elabore el diagrama de dispersión.
b) Determine ecuación de regresión con cualquiera de los 3 métodos presentados.
c) Grafique la recta de regresión en el diagrama de dispersión.
d) Verifique con Excel que los cálculos realizados sean correctos.

13.4 Modelo de regresión y sus supuestos

El modelo utilizado hasta aquí es:

$$\hat{y} = a + bx$$

Se basa en la forma pendiente-ordenada al origen de una recta. Representa un modelo determinístico, ya que dado un valor de x se obtiene uno para y. Tiene la siguiente forma general:

$$y = \beta_0 + \beta_1 x$$

El modelo probabilístico que incluye el componente de error es:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

En donde ε es el término del error, precisamente la diferencia entre cada uno de los valores observados de la variable y; y los estimados para cada valor de x, de acuerdo con la ecuación de regresión de mínimos cuadrados, como se vio en el ejemplo anterior. Este modelo se basa en 5 importantes suposiciones:

1. x es una variable aleatoria.
   La media de la distribución de probabilidad de ε es cero, lo cual implica que el valor esperado del error es cero, es decir, $E(\varepsilon) = \sum(y_1 - \hat{y}_i) = 0$
2. La distribución de probabilidad de ε tiene forma normal.
3. Existe una subpoblación de valores y para cada valor de x, esas subpoblaciones se distribuyen de forma normal.
4. Todas las varianzas de las subpoblaciones de y tienen la misma varianza.
5. Todas las medias de las subpoblaciones de y caen sobre la recta de regresión.

Existen métodos para probar estadísticamente estos supuestos pero rebasan el alcance de este libro. Aquí solamente se ilustra que el supuesto 2 permite hacer una comprobación empírica, ya que la suma de todos los errores, cuando se construye una ecuación de regresión lineal, debe ser igual a cero. De vuelta al ejemplo de los ingresos y el ahorro, en la tabla 13.5 se reproducen los datos. En la cuarta columna se calculan los valores de $\hat{y}$ para los valores dados de la variable independiente x. En la quinta columna se calcularon los errores $\varepsilon = y - \hat{y}$.

Tabla 13.5 Cálculos para ilustrar que $\sum \varepsilon = \sum (y_i - \hat{y}_i) = 0$

	Ingreso	Ahorro		
	x	y	$\hat{y} = 0.5367 + 0.1149x$	$\varepsilon = y - \hat{y}$
1	11	0.5	0.7272	–0.2272
2	14	1.1	1.0719	0.0281
3	12	0.9	0.8421	0.0579
4	9	0.6	0.4974	0.1026
5	13	1.2	0.957	0.243
6	13	0.9	0.957	–0.057
7	15	1.5	1.1868	0.3132
8	17	1.3	1.4166	–0.1166
9	15	1.1	1.1868	–0.0868
10	13	0.7	0.957	–0.257
			Suma	0.0002

En la suma de los errores existe una pequeña diferencia que se debe al redondeo, aún así puede apreciarse que el resultado es prácticamente cero.

ejercicios 13.4 Los supuestos del modelo de regresión lineal

Para los 5 ejercicios de la sección 13.3, verifique ahora que se cumple el supuesto 2 del modelo:

$$E(\varepsilon) = \sum (y_i - \hat{y}_i) = 0$$

13.5 Sumas de cuadrados en el análisis de regresión

Como se recordará, el método de los mínimos cuadrados para determinar la ecuación de regresión lineal garantiza que es mínima la suma de los cuadrados de las desviaciones entre cada observación y los valores de y estimados con esa recta de regresión, $\sum (y_i - \hat{y}_i)^2$. Esas diferencias entre los valores observados, y_i y los valores estimados mediante la ecuación de regresión, $\hat{y}_i$, pueden dividirse en 2 componentes con respecto al promedio de los valores observados, $\overline{y}$, tal como se ilustra en la figura 13.6.

En el diagrama se aprecia que dado un valor de x denotado $\hat{x}$, la desviación total de y con respecto a la media $\overline{Y}$, se divide en 2 partes:

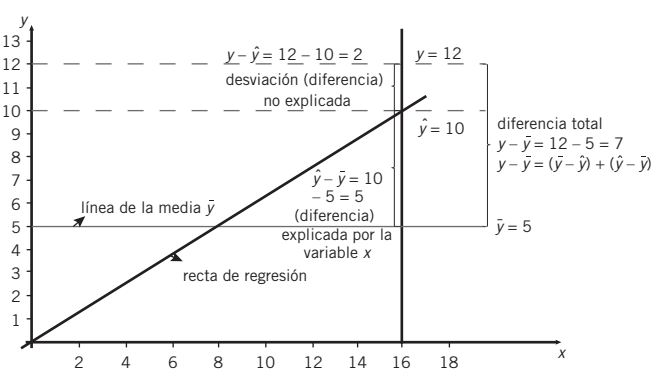


Figura 13.6 Diferencias $(y - \overline{y}) = (y - \hat{y}) + (\hat{y} - \overline{y})$.

Desviación total = desviación no explicada + desviación explicada.

En símbolos:

$$(y - \overline{y}) = (y - \hat{y}) + (\hat{y} - \overline{y})$$

Por su parte la $\overline{y}$, la media aritmética de los valores de $\overline{y} = \frac{\sum y}{n}$, se obtiene sin que en los cálculos participen los valores de x. La $\hat{y}$ que representa la recta de regresión, $y = a + bx$, se obtiene con la influencia de los valores de x. Si los valores de y están relacionados con los valores de x en algún grado, las desviaciones de los valores de $\hat{y}$ con respecto a $\overline{y}$ deben reducirse debido a la inclusión de los valores de x al calcular los valores de $\hat{y}$.

En otras palabras, la variación explicada representa la parte de la desviación total resultado de la inclusión de los valores de x en el cálculo de $\overline{y}$. La desviación explicada $(\hat{y} - \overline{y})$ es afectada o reducida por el uso de la variable x. A esta desviación explicada se le conoce también como desviación de regresión.

Por otro lado, la desviación no explicada $(y - \hat{y})$ es la que se mantiene a pesar de la introducción de los valores x al calcular la recta de regresión. Por eso, a esta desviación no explicada se le conoce también como desviación del error.

En resumen:

Desviación total = desviación del error + desviación de regresión. Desviación total = desviación no explicada + desviación explicada.

$$(y - \bar{y}) = (y - \hat{y}) + (\hat{y} - \bar{y})$$

Ahora, si se suman los cuadrados de estas diferencias para todos los puntos se tiene:

$$\sum (y_i - \bar{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2 \tag{13.8}$$

Suma de cuadrados total = suma de cuadrados del error + suma de cuadrados de regresión:

$$SCT = SCE + SCR \tag{13.9}$$

En el ejemplo siguiente se ilustra numéricamente la relación entre los 3 tipos de sumas de cuadrados.

EJEMPLO 13.4

Calcule las diferentes sumas de cuadrados para la ecuación y recta de regresión determinados para los datos de ingreso y ahorro.

Solución: En la tabla 13.6 se resumen los cálculos.

Tabla 13.6 Cálculos para las sumas de cuadrados de los datos de ingreso y ahorro.

	Ingreso	Ahorro		D total	D de regresión	D del error
	x	y	$\bar{y}$	$(y_i - \bar{y})^2$	$(\hat{y}_i - \bar{y})^2$	$(y_i - \hat{y}_i)^2$
1	11	0.5	0.7272	0.2304	0.0639	0.0516
2	14	1.1	1.0719	0.0144	0.0084	0.0008
3	12	0.9	0.8421	0.0064	0.0190	0.0034
4	9	0.6	0.4974	0.1444	0.2329	0.0105
5	13	1.2	0.957	0.0484	0.0005	0.0590
6	13	0.9	0.957	0.0064	0.0005	0.0032
7	15	1.5	1.1868	0.2704	0.0428	0.0981
8	17	1.3	1.4166	0.1024	0.1906	0.0136
9	15	1.1	1.1868	0.0144	0.0428	0.0075
10	13	0.7	0.957	0.0784	0.0005	0.0660
Sumas	132	9.8		0.916	0.6020	0.3139

Puede comprobarse a partir de los datos de la tabla:

$$\sum (y_i - \bar{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2$$

$$SCT = SCE + SCR$$

$$0.916 = 0.3139 + 0.602$$

Con una diferencia de una diezmilésima debida al redondeo.

Por supuesto, haciendo los despejes pertinentes pueden establecerse diferentes relaciones entre estas 3 sumas de cuadrados:

$$SCR = SCT - SCE$$

$$SCE = SCT - SCR$$

Las relaciones entre las 3 sumas de cuadrados se utilizan para 3 propósitos:

1. Evaluar la adecuación del modelo.
2. Calcular diferentes medidas que se utilizan tanto en el análisis de correlación como en la evaluación de la dispersión de los datos alrededor de la recta de regresión.
3. En los procedimientos que se siguen para hacer pruebas de hipótesis y estimaciones por intervalo de la pendiente de la recta.

En las secciones siguientes se analizarán estos temas; en seguida se revisa el importante concepto de la *desviación estándar de regresión*.

13.6 Desviación estándar de regresión

Desviación estándar de regresión. Es la desviación estándar de los valores y con respecto a los valores $\hat{y}$.

La **desviación estándar de regresión** es la desviación estándar de los valores y con respecto a los valores $\hat{y}$; es representada mediante el símbolo s, y es un estimador del verdadero valor de la población, σ. La fórmula que resume el procedimiento para calcularla es:

$$s = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{SCE}{n-2}} \tag{13.10}$$

En esta fórmula se aprecia la raíz cuadrada del promedio de los cuadrados de las desviaciones entre cada valor observado de y y su correspondiente $\hat{y}$, calculado mediante la ecuación de regresión. Por ello, mientras más cercanos estén los puntos de la recta de regresión menores serán esas diferencias (y sus cuadrados), consecuentemente menor será esta desviación estándar de regresión. Como ésta puede utilizarse para hacer estimaciones por intervalo de valores y, cuanto menor sea s, más precisas serán las estimaciones, ya que los intervalos serán más cortos. Y a la inversa, mientras mayores sean las distancias entre los valores y y sus correspondientes valores $\hat{y}$, mayor será la desviación estándar de regresión y menos precisas serán las estimaciones de y, dado que los intervalos serán más amplios.

Ejemplo 13.5

Calcule la desviación estándar de regresión para los datos de ingresos y ahorro.

Solución: Se reproducen los datos de ahorro e ingresos en la tabla 13.7, así como los valores $\hat{y}$, calculados en la tabla 13.5. Además se incluyen las operaciones necesarias para calcular la desviación estándar de regresión.

Tabla 13.7 Datos y operaciones para calcular $s_{y\hat{y}}$, la desviación estándar de regresión.

Ingreso	Ahorro		
x	y	$\hat{y}$	$(y_i - \hat{y}_i)^2$
11	0.5	0.7272	0.0516
14	1.1	1.0719	0.0008
12	0.9	0.8421	0.0034
9	0.6	0.4974	0.0105
13	1.2	0.957	0.0590
13	0.9	0.957	0.0032
15	1.5	1.1868	0.0981
17	1.3	1.4166	0.0136
15	1.1	1.1868	0.0075
13	0.7	0.957	0.0660
		Suma	0.3139

Así, la desviación estándar de regresión es:

$$s = \sqrt{\frac{\sum (y_i - \hat{y}_i)2}{n-2}} = \sqrt{\frac{SCE}{n-2}} = \sqrt{\frac{0.3139}{10-2}} = \sqrt{0.0392375} = 0.1981$$

13.7 Inferencias estadísticas sobre la pendiente β_1

Una inferencia importante sobre la ecuación de regresión es la prueba de hipótesis que realiza para evaluar si la pendiente de la recta, β_1, es diferente de cero, de ser así significa que no hay relación. En otras palabras, no hay pendiente (inclinación) y se trata de una ecuación que representa una recta horizontal.

Con esa prueba sobre la pendiente, cuando es diferente de cero pueden elaborarse estimaciones por intervalo sobre su valor. En seguida se revisan los procedimientos a seguir para hacer ambas inferencias.

13.7.1 Pruebas de hipótesis sobre la pendiente β_1

Para realizar inferencias (pruebas de hipótesis o estimaciones de intervalo) sobre este parámetro se requiere del error estándar correspondiente. En este caso, si se cumplen las suposiciones del modelo, la distribución muestral de β_1 es normal y su error estándar (la desviación estándar de la distribución muestral) se define como:

$$\sigma_{\hat{\beta}_1} = \frac{\sigma}{\sqrt{SC_{xx}}} \tag{13.11}$$

en donde σ es la desviación estándar de regresión que se vio en la sección anterior, su estimador es $s_{y\hat{y}}$, y SC_{xx} es la suma de los cuadrados de las diferencias entre cada una de las observaciones de la variable x y su promedio, $\bar{x}$ (o sea la suma de cuadrados que dividida entre n produce la varianza de x) la cual se utilizó para determinar los parámetros de la ecuación de regresión de mínimos cuadrados en la sección 13.3.3.

Ejemplo 13.6

Nuevamente con los datos de ingreso y ahorro, en el ejemplo 13.5 se encontró que la desviación estándar de regresión de esos datos es $s = 0.1981$. Se detectó también en la sección 13.3.3 que:

$$SC_{xx} = 45.6$$

de donde:

$$s_{\hat{\beta}_1} = \frac{s}{\sqrt{SC_{xx}}} = \frac{0.1981}{\sqrt{45.6}} = \frac{0.1981}{6.7528} = 0.02934$$

Por lo general se desconoce el verdadero valor de σ, así que suele utilizarse s como un estimador:

$$s_{\hat{\beta}_1} = \frac{s}{\sqrt{SC_{xx}}} \qquad \textbf{(13.12)}$$

La hipótesis que se quiere probar es que no existe relación entre las 2 variables, lo que equivale a decir que la pendiente de la recta de regresión es igual a cero. Esto contra una hipótesis alternativa de que la pendiente es diferente de cero, para una prueba de 2 extremos. En símbolos:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

Existen 2 maneras de probar esta hipótesis nula, a través de la t de Student y utilizando la F de Fisher. Se ilustran en seguida ambos casos.

El estadístico de prueba para la t de Student es:

$$t = \frac{\hat{\beta}_1 - \beta_1}{s\beta_1} \qquad \textbf{(13.13)}$$

El estadístico de prueba para la F de Fisher es:

$$F = \frac{\frac{SCR}{1}}{\frac{SCE}{n-2}} = \frac{SCR(n-2)}{SCE} \qquad \textbf{(13.14)}$$

13.7.1.1 Prueba sobre la pendiente β_1 utilizando la t de Student

Con los datos de ingreso y ahorro y utilizando un nivel de significación de 0.01 se tiene que:

Las hipótesis:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

El estadístico de prueba: $t = \dfrac{\hat{\beta}_1 - \beta_1}{s_{\beta_1}} = \dfrac{0.1149 = 0}{0.02934} = 3.916$

Dado el valor crítico de la t para una prueba de 2 extremos, $\alpha = 0.01$ y $n - 2$ grados de libertad (los mismos asociados con s), se tiene que:

$$P\left(-3.355 \leq t \leq 3.355\right) = 0.99$$

Se rechaza la hipótesis nula para concluir que sí existe una relación entre las 2 variables, en otras palabras, β_1 es diferente de cero.

13.7.1.2 Prueba sobre la pendiente utilizando la F de Fisher

Las hipótesis:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

El estadístico de prueba:

$$F = \frac{\frac{SCR}{1}}{\frac{SCE}{n-2}} = \frac{SCR(n-2)}{SCE}$$

Con 1 grado de libertad para el numerador y $n - 2$ para el denominador. Si se había fijado un nivel de significación de 0.01 se tiene que el valor crítico de la F, de acuerdo con la tabla 4 del apéndice, es:

$$P(F \geq 11.26 \mid gl_{num} = 1,\ gl_{denom} = 8) = 0.01$$

El valor calculado con las cifras encontradas antes es:

$$F = \frac{\frac{SCR}{1}}{\frac{SCE}{n-2}} = \frac{SCR(n-2)}{SCE} = \frac{0.6020(10-2)}{0.3139} = \frac{4.816}{0.3139} = 15.34$$

Debido a que el valor calculado de F es mayor que su valor crítico, se rechaza la hipótesis nula y se concluye que la pendiente de la ecuación de regresión no es cero y que sí existe una regresión de y sobre x. El que se haya llegado a la misma conclusión mediante los 2 estadísticos de prueba, t y F, no es casualidad, ya que los 2 métodos son equivalentes.

13.7.2 Estimación por intervalo de β_1

Una estimación por intervalo de β_1 con un nivel de confianza de 99% se construye de la siguiente manera:

$$\hat{\beta} \pm ts_{\beta_1}$$

$$\hat{\beta} \pm ts_{\beta_1} = 0.1149 \pm 3.355(0.02934) = 0.1149 \pm 0.0984 \qquad \textbf{(13.15)}$$

Así, con una confianza de 99% de estar en lo correcto, se estima que la pendiente de la recta de regresión que relaciona el ingreso con el ahorro se encuentra entre 0.01609 y 0.2133.

EJERCICIOS 13.7 Inferencias estadísticas sobre la pendiente β_1

1. Retome los 5 ejercicios de la sección 13.3 y siga las instrucciones que se muestran a continuación.

a) Realice la prueba de hipótesis sobre β_1 para determinar si existe relación entre las 2 variables. Utilice $\alpha = 0.01$.

b) Haga una estimación por intervalo para β_1 con un nivel de confianza de 99 por ciento.

13.8 Uso de la ecuación de regresión para estimación y predicción

Una vez que se ha construido la ecuación de regresión, que se ha realizado la prueba de hipótesis sobre si es o no igual a cero su pendiente, y si se está satisfecho sobre el cumplimiento de los supuestos del modelo, se procederá a utilizar la ecuación en los 2 propósitos principales para los que fue construida, es decir, a hacer estimaciones o predicciones para la variable dependiente con base en valores de la variable independiente:

1. Estimar el valor promedio o esperado de y, E(y), para un valor específico de x.
2. Pronosticar sobre la variable y, con base en valores dados de x.

En el caso de la estimación de un valor promedio de y se intenta calcular el resultado promedio de un número grande de ensayos realizados sobre el valor dado de x; para el pronóstico se intenta predecir el resultado de un solo ensayo para el valor dado de x. En seguida se revisan los 2 casos.

13.8.1 Estimación por intervalo de *y* para valores dados de *x*

En el ejemplo de los ingresos y el ahorro podría utilizarse la ecuación de regresión para estimar cuánto ahorro podría esperarse en un nivel determinado de ingresos. En otro caso, si las utilidades de una empresa están relacionadas linealmente con el gasto en publicidad, se emplearía la ecuación de regresión construida con estas 2 variables. El objetivo sería estimar el grado de utilidades que se esperaran en un nivel determinado de gastos en publicidad. En otro ejemplo, si el tiempo dedicado a capacitación está linealmente relacionado con la productividad de los programadores de computadoras, se estimaría la cantidad de código que produjera un programador para un determinado tiempo dedicado a la capacitación.

Para hacer estas estimaciones se utiliza la desviación estándar de la distribución muestral de $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$, es decir, su error muestral que es:

$$\sigma_{\hat{y},e} = \sigma\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{SC_{xx}}} \qquad \textbf{(13.16)}$$

El estadístico inferencial es la t de Student con $n-2$ grados de libertad, por lo que el intervalo se construye como:

$$\hat{y} \pm ts\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{SC_{xx}}} \tag{13.17}$$

Nótese que en esta última expresión, se cambió σ por s, debido a que los cálculos se realizan con valores muestrales.

Ejemplo 13.7

Para el ejemplo de los ingresos y el ahorro haga una estimación por intervalo para el valor de y, dado $x = 10$, con un nivel de confianza de 95 por ciento.

Solución: Se comprobó con anterioridad que la ecuación de regresión, obtenida con Excel es:

$$\hat{y} = -0.53684211 + 0.11491228x$$

El valor de y correspondiente a $x = 10$ es:

$$\hat{y} = -0.53684211 + 0.11491228(10) = -0.5368 + 1.1491 = 0.6123$$

La desviación estándar de regresión se calculó en el ejemplo 13.5:

$$s = 0.1981$$

El valor de t para una confianza de 95% y grados de libertad $n - 2 = 10 - 2 = 8$, es:

$$P(-2.306 \le t \le 2.306 | \text{g}\, l = g) = 0.95$$

Y:

$$\hat{y} \pm ts\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{SC_{xx}}} = 0.6123 \pm 2.306(0.1981)\sqrt{\frac{1}{10}+\frac{(10-13.2)^2}{45.6}}$$

$$= 0.6123 \pm 0.457\sqrt{0.1+0.225} = 0.6123 \pm 0.57 = 0.042 - 1.182$$

Se estima, con una confianza de 95% de estar en lo correcto, que el nivel de ahorro para una persona con un ingreso de 10 está entre 0.042 y 1.182, o dado que las cantidades están expresadas en miles de pesos, que el nivel de ahorro para una persona que tiene un ingreso de \$10 000 está entre \$42 y \$1 182.

13.8.2 Pronósticos de *y* para valores dados de *x*

La única diferencia en el procedimiento para construir intervalos de predicción con respecto a los intervalos de estimación es el error estándar. Para el caso de la predicción este error estándar es:

$$\sigma_{\hat{y},\,p} = \sigma\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{SC_{xx}}} \tag{13.18}$$

Dado que se utilizan valores muestrales para estimarlo se convierte en:

$$s_{\hat{y},p} = s\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{SC_{xx}}} \tag{13.19}$$

Por lo que el intervalo de pronóstico es:

$$\hat{y} \pm ts\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{SC_{xx}}} \tag{13.20}$$

EJEMPLO 13.8

Para el ejemplo de los ingresos y el ahorro haga un pronóstico para el valor de y, dado $x = 10$, con un nivel de confianza de 95 por ciento.

Solución: En el ejemplo 13.7 se encontró que el valor de y, a partir de la ecuación de regresión cuando $x = 10$, es 0.6123.

El intervalo de predicción:

$$\hat{y} \pm ts\sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{SC_{xx}}} - 0.6123 \pm 2.306(0.1981)$$

$$\sqrt{1 + \frac{1}{10} + \frac{(10 - 13.2)^2}{45.6}}$$

$$= 0.6123 \pm 0.4568\sqrt{1.1 + 0.2246} = 0.6123 \pm 0.4568(1.15)$$

$$= 0.6123 \pm 0.53$$

Por lo que el intervalo de pronóstico, con un nivel de confianza de 95%, es de 0.09 a 1.14.

EJERCICIOS 13.8 Estimación por intervalo y pronósticos de y para valores dados de x

1. Retome el ejercicio 1 de los ejercicios 13.3 y siga las instrucciones que se muestran a continuación:

a) Estime, con un nivel de confianza de 99%, el tiempo de entrega para un embarque que va a una distancia de 1 500 km.

b) Pronostique, con un nivel de confianza de 99%, el tiempo de entrega para un embarque que va a una distancia de 1 500 km.

2. Retome el ejercicio número 2 de los ejercicios 13.3 y siga las instrucciones que se muestran a continuación:

a) Estime, con un nivel de confianza de 95%, la producción de caña esperada con 12 kg de fertilizante.

b) Pronostique, con un nivel de confianza de 95%, la producción de caña con 12 kg de fertilizante.

3. Retome el ejercicio número 3 de los ejercicios 13.3 y siga las instrucciones que se muestran a continuación:

a) Estime, con un nivel de confianza de 99%, la cantidad de productos rechazados elaborados por un trabajador con 12 semanas de experiencia.

b) Pronostique, con un nivel de confianza de 99%, la cantidad de productos rechazados elaborados por un trabajador con 12 semanas de experiencia.

4. Retome el ejercicio número 4 de los ejercicios 13.3 y siga las instrucciones que se muestran a continuación:

a) Estime, con un nivel de confianza de 95%, la calificación de un alumno que dedica 6 horas al estudio.

b) Pronostique, con un nivel de confianza de 95%, la calificación de un alumno que dedica 6 horas al estudio.

5. Retome el ejercicio número 5 de los ejercicios 13.3 y siga las instrucciones que se muestran a continuación:

a) Estime, con un nivel de confianza de 99%, el volumen de ventas de un vendedor con 11 años de experiencia.

b) Pronostique, con un nivel de confianza de 99%, el volumen de ventas de un vendedor con 11 años de experiencia.

13.9 Recapitulación del análisis de regresión lineal simple

Vale la pena hacer un recuento de lo revisado hasta aquí, ya que ayudará a la comprensión del tema y a la aplicación de lo estudiado.

El punto de partida es que el análisis de regresión lineal simple se refiere al estudio de la relación lineal, es decir, rectilínea entre 2 variables. Se analiza si una nube de puntos que representa los valores pareados de esas 2 variables tiene apariencia de ajustarse a una línea recta. Si es así, se ajusta una recta a los puntos; el ajuste puede hacerse de forma manual (aunque el procedimiento es inexacto) o mediante la técnica conocida como *mínimos cuadrados*, la cual garantiza que las distancias verticales entre cada uno de los puntos y la recta de regresión sean mínimas.

El procedimiento de mínimos cuadrados se basa en la resolución simultánea de 2 ecuaciones, las ecuaciones normales. Una vez obtenida la recta y su ecuación, se necesita evaluar si el modelo representa realmente una relación lineal. Así que se hace una prueba de hipótesis sobre la pendiente de la recta para poder asumirla diferente de cero. En caso de que sea igual a cero se tiene una recta horizontal lo que significa que no existe relación y el análisis se detendría.

Si, por el contrario, se prueba que la pendiente es diferente de cero y se asume que se cumplen los supuestos en los que se basa el modelo, entonces se utiliza en la práctica esa ecuación de regresión para

hacer estimaciones sobre el posible valor de la variable dependiente para un determinado valor de la variable independiente.

En todo este proceso se hicieron varias precisiones: se comienza por hacer un ajuste a mano alzada sobre la nube de puntos para dibujar una recta que se le ajuste; a partir de ésta, se construye la ecuación de regresión. A pesar de que este método resulta ilustrativo es considerablemente impreciso, así que se introduce el método de los mínimos cuadrados, el mejor para construir la ecuación que describe a la recta con mejor ajuste a la nube de puntos. Se trata de un método que se basa en la resolución simultánea de un par de ecuaciones conocidas como *ecuaciones normales.*

Se revisaron 2 métodos para derivar estas ecuaciones normales, uno puramente algebraico y otro basado en cálculo diferencial que permite asegurar que se trata, precisamente, de la ecuación de mínimos cuadrados.

También se explicaron 4 formas para resolver las ecuaciones normales que encuentren los parámetros de la ecuación de regresión, su pendiente y su ordenada al origen. Esos 4 métodos son:

1. Despeje simultáneo de a y b en las 2 ecuaciones normales.
2. Resolución simultánea de las 2 ecuaciones normales.
3. Uso de promedios y de sumas de cuadrados.
4. Uso de Excel.

De estos 4 métodos, el más fácil es el mecanismo que provee Excel, ya que arroja en un santiamén toda la información requerida. Cabe resaltar que el uso de promedios y de sumas de cuadrados tiene varias ventajas, especialmente útiles para quienes comienzan a estudiar estos temas. Una de éstas es que permite seguir de cerca lo que se está haciendo y se comprenden mejor los resultados con sus posibles aplicaciones. Por esta razón se resume desde este punto de vista lo que se hizo en esta primera mitad del capítulo con base en el ejemplo de la relación entre los ingresos y los ahorros de un grupo de personas. Se comenzó con el ejemplo.

EJEMPLO 13.9

El gerente de un banco desea saber si puede considerarse que el ahorro de las familias (variable y) depende de sus ingresos (variable x). En la tabla 13.8 se muestran los resultados que se obtienen para una muestra de 10 familias.

Tabla 13.8 Ingresos y ahorros de 10 familias (en miles de pesos mensuales).

Familia	Ingresos (x)	Ahorro (y)
1	11	0.5
2	14	1.1
3	12	0.9
4	9	0.6
5	13	1.2
6	13	0.9
7	15	1.5
8	17	1.3
9	15	1.1
10	13	0.7

Se graficó el diagrama de dispersión y se observó que las 2 variables parecían tener una relación rectilínea. Para derivar la ecuación de regresión utilizando promedios y sumas de cuadrados, así como ecuaciones normales:

$$\sum y = na + b\sum x \qquad \textbf{(I)}$$

$$\sum xy = a\sum x + b\sum x^2 \qquad \textbf{(II)}$$

Se utilizaron las ecuaciones (13.6) y (13.7) que se basan precisamente en sumas de cuadrados:

$$a = \bar{y} - b\bar{x} \qquad \textbf{(13.5)}$$

$$b = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2} = \frac{SC_{xy}}{SC_{xx}} \qquad \textbf{(13.7)}$$

Entonces, se completó la tabla de datos con las operaciones necesarias para encontrar estos promedios y sumas de cuadrados. Se llegó a la tabla 13.9:

Tabla 13.9 Datos y operaciones para resolver el ejercicio de ingreso y ahorro utilizando sumas de cuadrados

	Ingreso	Ahorro		
	x	y	$(x-\bar{x})^2$	$(x-\bar{x})(y-\bar{y})$
1	11	0.5	4.84	1.056
2	14	1.1	0.64	0.096
3	12	0.9	1.44	0.096
4	9	0.6	17.64	1.596
5	13	1.2	0.04	–0.044
6	13	0.9	0.04	0.016
7	15	1.5	3.24	0.936

	Ingreso	Ahorro		
	x	y	$(x-\bar{x})^2$	$(x-\bar{x})(y-\bar{y})$
8	17	1.3	14.44	1.216
9	15	1.1	3.24	0.216
10	13	0.7	0.04	0.056
Sumas	132	9.8	45.6	5.24

Con estos datos se construyó la ecuación de regresión:

$$b = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2} = \frac{SC_{xy}}{SC_{xx}} = \frac{5.24}{45.6} = 0.1149$$

Y:

$$a = \bar{y} - b\bar{x} = \frac{9.8}{10} - 0.1149\frac{132}{10} = 0.98 - 0.1149(13.2) = -0.5367$$

Se llegó a la siguiente ecuación de regresión:

$$\hat{y} = -0.5367 + 0.1149x$$

Posteriormente se utilizó una relación entre las sumas de cuadrados de las desviaciones de regresión, ilustrada en la figura 13.6, que arranca a partir de las desviaciones de cualquier punto de datos y valores estimados provenientes de la ecuación de regresión; resumidos como $(y-\bar{y}) = (y-\hat{y}) + (\hat{y}-\bar{y})$:

Desviación total = desviación no explicada + desviación explicada. Cuyas sumas de cuadrados son:

$$\sum(y_i-\bar{y})^2 = \sum(y_i-\hat{y}_i)^2 + \sum(\hat{y}_i-\bar{y})^2$$

$$\text{SCT} \quad = \quad \text{SCE} \quad + \quad \text{SCR}$$

En resumen, suma de cuadrados total (SCT) = suma de cuadrados del error (SCE) + suma de cuadrados de regresión (SCR).

Esta equivalencia de sumas de cuadrados permite, por un lado, verificar que los cálculos se han realizado en forma apropiada —si al revisar esos cálculos no se cumple esta igualdad entonces se sabe que algo no se hizo bien— y calcular otros parámetros importantes para las inferencias estadísticas relacionadas con el análisis de regresión. Ahora se combina la tabla 13.3 con la 13.6 para reunir todas las sumas de cuadrados que se utilizan en la tabla 13.10.

Tabla 13.10 Resumen de las sumas de cuadrados utilizadas en el análisis de regresión lineal simple

Ingreso x	Ahorro y	Cuadrados de diferencias $(x-\bar{x})^2$	Diferencias x por diferencias y $(x-\bar{x})(y-\bar{y})$	Cuadrados de diferencias $(y_i-\bar{y})^2$	Cuadrados de diferencias $(\hat{y}_i-\bar{y})^2$	Cuadrados de diferencias $(y_i-\hat{y}_i)^2$
11	0.5	4.84	1.056	0.2304	0.0639	0.0516
14	1.1	0.64	0.096	0.0144	0.0084	0.0008
12	0.9	1.44	0.096	0.0064	0.0190	0.0034
9	0.6	17.64	1.596	0.1444	0.2329	0.0105
13	1.2	0.04	–0.044	0.0484	0.0005	0.0590
13	0.9	0.04	0.016	0.0064	0.0005	0.0032
15	1.5	3.24	0.936	0.2704	0.0428	0.0981
17	1.3	14.44	1.216	0.1024	0.1906	0.0136
15	1.1	3.24	0.216	0.0144	0.0428	0.0075
13	0.7	0.04	0.056	0.0784	0.0005	0.0660
132	9.8	45.6	5.24	0.916	0.6020	0.3139
Suma de x	Suma de y	Suma de cuadrados de diferencias $(x-\bar{x})^2$	Suma de diferencias x por diferencias y $(x-\bar{x})(y-\bar{y})$	Suma de cuadrados de diferencias $(y_i-\bar{y})^2$	Suma de cuadrados de diferencias $(\hat{y}_i-\bar{y})^2$	Suma de cuadrados de diferencias $(y_i-\hat{y}_i)^2$
		SC_{xx}	SC_{xy}	SCT SC_{yy}	SCR	SCE

Con ayuda de estas sumas de cuadrados se calculó la desviación estándar de regresión:

$$s = \sqrt{\frac{\sum(y_i-\hat{y}_i)^2}{n-2}} = \sqrt{\frac{SCE}{n-2}} = \sqrt{\frac{0.3139}{10-2}} = \sqrt{0.0392375} = 0.1981$$

Con esta desviación estándar de regresión se calculó el error estándar (la desviación estándar de la distribución muestral) de la pendiente de la recta de regresión de la población β_1:

$$\sigma_{\hat{\beta}_1} = \frac{\sigma}{\sqrt{SC_{xx}}}$$

Se estimó utilizando el valor muestral s:

$$s_{\beta_1} = \frac{\sigma}{\sqrt{SC_{xx}}} = \frac{0.1981}{\sqrt{45.6}} = \frac{0.1981}{6.7528} = 0.02934$$

la cual se utilizó para probar la hipótesis: $H_0 : \beta_1 = 0$

Se utilizó como estadístico de prueba la t de Student:

$$t = \frac{\hat{\beta}_1 - \beta_1}{s\beta_1} \qquad \textbf{(13.13)}$$

Por otro lado, se probó también la hipótesis: $H_0 : \beta_1 = 0$

Como estadístico de prueba se empleó la F de Fisher:

$$F = \frac{\frac{SCR}{1}}{\frac{SCE}{n-2}} = \frac{SCR(n-2)}{SCE}$$

Nuevamente se utilizaron sumas de cuadrados.

Con ambas pruebas se concluyó que sí existía relación entre las 2 variables, ya que la pendiente resultó ser diferente de cero.

Como parte final de las inferencias sobre la pendiente de la recta, utilizando varios de los resultados anteriores se construyó un intervalo de confianza para esa pendiente:

$$\hat{\beta} \pm ts_{\beta_1} = 0.1149 \pm 3.355(0.02934) = 0.1149 \pm 0.0984$$

Así se concluyó el análisis de regresión utilizando la ecuación para:

a) Hacer estimaciones por intervalo de la variable dependiente, y, para un valor dado de x, para lo cual se requirió calcular también el error muestral de la recta de regresión, $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$, que es:

$$\sigma_{\hat{y}} = \sigma\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{SC_{xx}}}$$

En donde vuelven a aparecer sumas de cuadrados, implícitas como en σ, y explícitas en SC_{xx}. Al contar con este error estándar se construye el intervalo de estimación:

$$\hat{y} \pm ts_{\hat{y},e}\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{SC_{xx}}} \qquad \textbf{(13.17)}$$

b) Hacer pronósticos sobre el valor que se esperaría obtener para y con un valor determinado para x, para lo cual el error estándar correspondiente es:

$$S_{\hat{y},p} = s\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{SC_{xx}}}$$

El correspondiente intervalo es:

$$\hat{y} \pm ts\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{SC_{xx}}}$$

Ahora se resolverá un ejercicio con esta manera de utilizar las diferentes sumas de cuadrados.

EJEMPLO 13.10

El 29 de abril de 2011 se publicó en el periódico *La Jornada* un artículo del economista Julio Boltvinik en el que propone una nueva metodología para medir la incidencia de la pobreza a nivel nacional a través de los costos de los bienes familiares variables. Cita los siguientes datos provenientes de una tesis de licenciatura de Alejandro Marín. Esta información puede consultarse en el siguiente sitio web: http://www.jornada.unam.mx/2011/04/29/index.php?section=economia&article=032o1eco

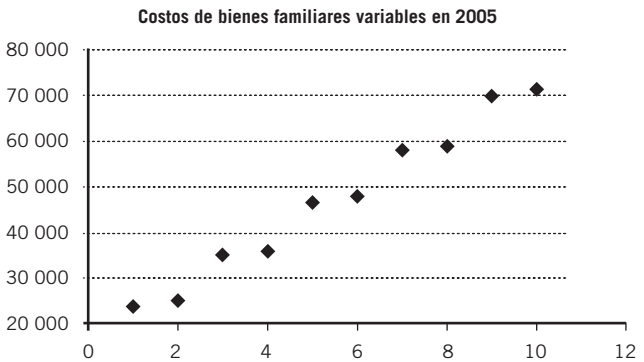
Personas por hogar x	Costos de bienes familiares variables en 2005 y
1	23 693
2	25 080
3	35 053
4	35 868
5	46 532
6	47 940
7	58 066
8	58 829
9	69 904
10	71 418

Con estos datos siga las instrucciones enlistadas a continuación:

a) Grafique el diagrama de dispersión de los puntos que representan estos datos.
b) Calcule la ecuación de regresión por el método de mínimos cuadrados.
c) Grafique la ecuación de regresión sobre la nube de puntos.
d) Pruebe la hipótesis nula de que la pendiente de esa recta es igual a cero.
e) Haga una estimación por intervalo, con 95% de confianza, de la pendiente de la recta de regresión.
f) Haga una estimación y un pronóstico por intervalo del costo de los bienes familiares variables en 2005 para una familia compuesta por 12 miembros, con un nivel de confianza de 95 por ciento.
g) Utilice la regresión del "Análisis de datos" de Excel, para verificar que los cálculos sean correctos.

Solución:

a) Gráfica del diagrama de dispersión de los puntos.



b) Las operaciones, en la tabla 13.11.

Tabla 13.11 Operaciones para calcular la ecuación de regresión para el ejemplo 13.9

Personas por hogar x	Costos de bienes familiares variables en 2005 y	Cuadrados de diferencias $(x-\bar{x})^2$	Diferencias x por diferencias y $(x-\bar{x})(y-\bar{y})$	Cuadrados de diferencias $(y_i-\bar{y})^2$
1	23693	20.25	105953.85	554381152.09
2	25080	12.25	77554.05	490990258.89
3	35053	6.25	30463.25	148481536.09
4	35868	2.25	17055.45	129283722.09
5	46532	0.25	353.15	498859.69
6	47940	0.25	350.85	492382.89
7	58066	2.25	16241.55	117239087.29
8	58829	6.25	28976.75	134344326.49
9	69904	12.25	79329.95	513733956.49
10	71418	20.25	108808.65	584657892.09
Σ = 55	Σ = 472383	82.50	465087.50	2674103174.10

$$b=\frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2}=\frac{SC_{xy}}{SC_{xx}}=\frac{465\,087.5}{82.5}=5\,637.42$$

Y:

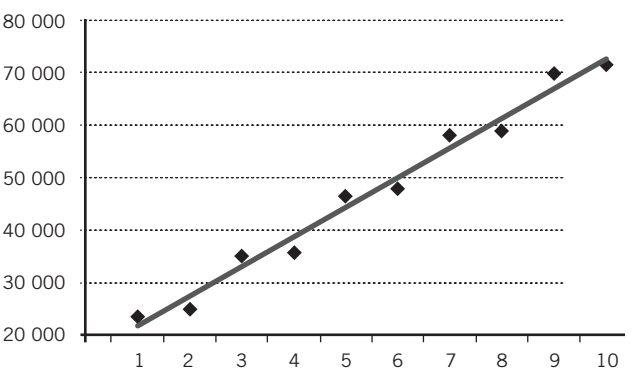
$$a=\bar{y}-b\bar{x}=\frac{472\,383}{10}-5\,637.42\,\frac{55}{10}=47\,238.30-5.637.42(5.5)$$

$$=16\,232.49$$

Por lo que la ecuación de regresión lineal es:

$$\hat{y}=16\,232.49+5\,637.42x$$

c) Gráfica de la ecuación de regresión sobre la nube de puntos.



d) En primer lugar, se calculan las sumas de cuadrados del error y de regresión, así se llega a la tabla 13.12:

Tabla 13.12 Todas las sumas de cuadrados para el ejemplo 13.9

Personas por hogar x	Costos de bienes familiares variables en 2005 y	Cuadrados de diferencias $(x-\bar{x})^2$	Diferencias x por diferencias y $(x-\bar{x})(y-\bar{y})$	Cuadrados de diferencias $(y_i-\bar{y})^2$	$\hat{y}$	Cuadrados de diferencias $(\hat{y}_i-\bar{y})^2$	Cuadrados de diferencias $(y_i-\bar{y}_i)^2$
1	23693	20.25	105953.85	554381152.09	21 869.91	643555211.19	3323657.15
2	25080	12.25	77554.05	490990258.89	27507.33	389311177.14	5891930.93
3	35053	6.25	30463.25	148481536.09	33144.75	198628151.60	3641418.06
4	35868	2.25	17055.45	129283722.09	38782.17	71506134.58	8492386.79
5	46532	0.25	353.15	498859.69	44419.59	7945126.06	4462276.01
6	47940	0.25	350.85	492382.89	50057.01	7945126.06	4481731.34
7	58066	2.25	16241.55	117239087.29	55694.43	71506134.58	5624344.26
8	58829	6.25	28976.75	134344326.49	61331.85	198628151.60	6264258.12
9	69904	12.25	79329.95	513733956.49	66969.27	389311177.14	8612640.17
10	71418	20.25	108808.65	584657892.09	72606.69	643555211.19	1412983.92
55	472383	82.50	465087.50	2674103174.10	472383.00	2621891601.15	52207626.75

t de Student como estadístico de prueba.

Después de realizar las sumas de cuadrados se calcula la desviación estándar de regresión:

$$s = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{SCE}{n-2}} = \sqrt{\frac{52\ 207\ 626.75}{10-2}} = 2\ 554.59$$

Con esta desviación estándar de regresión se calcula el error estándar de la pendiente de la recta de regresión de la población, β_1:

$$s_{\beta_1} = \frac{\sigma}{\sqrt{SC_{xx}}} = \frac{2\ 554.59}{\sqrt{82.5}} = \frac{18\ 282.86}{9.083} = 281.25$$

Luego de obtener este error estándar puede probarse la hipótesis sobre la pendiente de la recta:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

Si se utiliza como estadístico de prueba la t de Student:

$$t = \frac{\hat{\beta}_1 - \beta_1}{s_{\beta_1}} = \frac{5\ 637.42 - 0}{281.25} = 20.044$$

Ahora, el valor crítico de la t de Student para un nivel de significación de 0.05, $n - 2$ grados de libertad y una prueba de 2 extremos es:

$$P\left(-2.306 \leq t \leq 2.306 \middle| gl = 8\right) = 0.95$$

Debido a que el valor calculado del estadístico de prueba es superior a 2.306, se rechaza la hipótesis nula para concluir que la pendiente de la recta de regresión es diferente de cero, por lo tanto sí existe una relación lineal entre las 2 variables.

Con la F de Fisher como estadístico de prueba.

$$P\left(F \geq 5.32 \middle| gl_{num} = 1, gl_{denom} = 8\right) = 0.95$$

El valor calculado con las cifras encontradas antes es:

$$F = \frac{\frac{SCR}{1}}{\frac{SCE}{n-2}} = \frac{SCR(n-2)}{SCE} = \frac{2\ 621\ 891\ 601.15(10-2)}{52\ 207\ 626.75} = 401.76$$

Se observa que el valor calculado de F es mayor que su valor crítico, así que se rechaza la hipótesis nula y se concluye que la pendiente de la ecuación de regresión no es cero y sí existe una regresión de y sobre x.

e) La estimación por intervalo, con 95% de confianza, de la pendiente de la recta de regresión:

$$\hat{\beta} \pm ts_{\beta_1} = 5\ 637.42 \pm 2.306(281.25) = 5\ 306.47 \pm 648.56$$

Así se estima que la pendiente de la recta de regresión está entre 4 657.91 y 5 955.03.

f) La estimación por intervalo del costo de los bienes familiares variables en 2005 para una familia compuesta por 12 miembros, con un nivel de confianza de 95 por ciento.

Para una familia de 12 miembros:

$$\hat{y} = 16\ 232.49 + 5\ 637.42x = 16\ 232.49 + 5\ 637.42(12) = 83.881.53$$

Para una estimación, el error estándar de la recta de regresión, $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$:

$$s_{\hat{y}} = s\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{SC_{xx}}} = 2\ 554.59\sqrt{\frac{1}{10} + \frac{(12-5.5)^2}{82.5}}$$

$$= 2\ 554.59(0.78238) = 1\ 998.66$$

El intervalo es:

$$\hat{y} \pm ts_{\hat{y}}\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{SC_{XX}}} = 2\ 554.59\sqrt{\frac{1}{10} + \frac{(12-5.5)^2}{82.5}}$$

$$= 2\ 554.59(0.78238) = 1\ 998.66$$

Por lo que se estima, con una confianza de 95% de estar en lo correcto, que los costos de bienes familiares variables en 2005 para una familia de 12 miembros están entre \$79 051.95 y \$88 269.77.

Para un pronóstico, el error estándar de la recta de regresión, $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$:

$$s_{\hat{y}} = s\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{SC_{xx}}} = 2\ 554.59\sqrt{1 + \frac{1}{10} + \frac{(12-5.5)^2}{82.5}}$$

$$= 2\ 554.59(1.27) = 3\ 244.33$$

El intervalo es:

$$\hat{y} \pm ts_{\hat{y}}\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{SC_{XX}}} = 83\ 660.86 \pm 2.306(3\ 244.33)$$

$$= 83\ 660.86 \pm 7\ 481.42$$

Por lo que se pronostica, con una confianza de 95% de estar en lo correcto, que los costos de bienes familiares variables en 2005 para una familia de 12 miembros están entre \$76 179.44 y \$91 142.28.

g) La sección de "Análisis de varianza" de los resultados que se obtienen mediante la "Regresión" de Excel son:

Análisis de varianza						
	Grados de libertad	Suma de cuadrados	Promedio de los cuadrados	F	Valor crítico de F	
Regresión	1	2 621 895 547	2 621 895 547	401.7643721	4.0041E-08	
Residuos	8	52 207 627	6 525 953.344			
Total	9	2 674 103 174				

	Coeficientes	Error típico	Estadístico *t*	Probabilidad	Inferior 95%	Superior 95%
Intercepción	16 232.46667	1 745.120309	9.301631859	1.45337E-05	12 208.212	20 256.7213
x	5 637.424242	281.251604	20.04406077	4.00407E-08	4 988.85688	6 285.9916

En donde puede observarse que:

a) los coeficientes que aparecen al final son precisamente los coeficientes de la ecuación de regresión obtenidos previamente;
b) las sumas de cuadrados también son prácticamente las mismas que se muestran en la tabla 13.12.

Además, se aprecia que en el renglón de *regresión* se encuentra la misma *F* calculada con anterioridad en *d*): 401.76. En ese mismo renglón, el valor que aparece debajo de "Valor crítico de F", 4.0041E-08 (en notación científica y que equivale a 0.000000040041 en notación decimal) indica la probabilidad de haber obtenido ese valor de la *F* que es prácticamente de cero.

En la siguiente sección se revisará el tema de la correlación lineal que se ocupa de evaluar el sentido y la intensidad de la relación entre las 2 variables, expresada por la ecuación de regresión.

EJERCICIOS 13.9 Recapitulación del análisis de regresión lineal simple

1. Retome los 5 ejercicios de la sección 13.3 y siga las instrucciones que se muestran a continuación:

a) Grafique el diagrama de dispersión de los puntos que representan estos datos.
b) Calcule la ecuación de regresión por el método de mínimos cuadrados.
c) Grafique la ecuación de regresión sobre la nube de puntos.
d) Pruebe la hipótesis nula de que la pendiente de esa recta es igual a cero.
e) Haga una estimación por intervalo, con 95% de confianza, de la pendiente de la recta de regresión.
f) Utilice la "Regresión" del "Análisis de datos" de Excel para verificar que los cálculos sean correctos.
g) Haga por intervalo una estimación y pronóstico de *y*.

13.10 Análisis de correlación

El análisis de regresión y el de correlación están estrechamente ligados. Se ha revisado cómo el análisis de regresión lineal simple se utiliza para calcular la ecuación de una recta que minimiza los cuadrados de las desviaciones (verticales) de cada observación con respecto a la línea ajustada: $(y-\hat{y})^2$. También se ha identificado cómo se utiliza esta ecuación de regresión para estimar y pronosticar otros valores de la variable dependiente, *y*, para ciertos valores de la variable independiente, *x*. El análisis de regresión más sencillo es el que corresponde a:

- Una relación lineal, es decir, en forma de línea recta,
- Una relación simple, es decir, una relación entre 2 variables, *x* y *y*, la variable independiente y la dependiente, respectivamente.

Además del análisis de regresión puede aplicarse el *análisis de correlación* en el que se revisa, ya no el tipo de relación, sino la intensidad y el sentido de la relación entre las 2 variables. Esta intensidad de la relación suele medirse a través de 2 vías: el coeficiente de determinación que se denota por r^2 y el coeficiente de correlación denotado por *r*, el cual, tal como puede apreciarse en la simbología, es la raíz cuadrada de r^2, o sea $\sqrt{r^2} = r$.

El **coeficiente de determinación** r^2 se define como la razón de la variación explicada a la variación total, lo cual puede visualizarse mejor revisando la figura 13.6. Como puede apreciarse ahí, la variación explicada es $\hat{y}-\bar{y}$, en tanto que la variación total es $y_i-\bar{y}$, por lo que el coeficiente de determinación está dado por el cociente entre las sumas de los cuadrados de estas 2 desviaciones:

Coeficiente de determinación r^2. Es la razón de la variación explicada a la variación total.

$$r^2 = \frac{\text{Variación explicada}}{\text{Variación total}} = \frac{\sum(\hat{y}_i-\bar{y})^2}{\sum(y_i-\bar{y})^2} = \frac{SCR}{SCT}$$

O simplemente:

$$r^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = \frac{SCR}{SCT} \tag{13.21}$$

Analizando esta expresión puede entenderse que una correlación es perfecta cuando no hay dispersión alrededor de la recta de regresión, es decir, cuando todos los valores y_i caen sobre la línea de regresión $y_i = \hat{y}_i$, y se tiene que $\sum(\hat{y}_i - \bar{y})^2 = \sum(y_i - \bar{y})^2$,

$$r^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = \frac{\sum(y_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = 1$$

En otras palabras, no hay desviaciones entre los valores observados de y y los valores estimados mediante la recta de regresión. O sea el ajuste entra la nube de puntos y la recta de regresión es perfecta.

Por otra parte, cuando los puntos y_i están muy distanciados de la recta de regresión, $\sum(y_i - \hat{y}_i)^2$ se vuelve muy grande, y como la variación total es fija, $\sum(\hat{y}_i - \bar{y})^2$ se vuelve muy pequeña y el valor de r^2 se aproxima a cero, lo cual quiere decir que no existe correlación entre los puntos y la recta de regresión. Una conclusión importante que se desprende de lo anterior es que r^2 asume valores entre 0 y 1.

Expresado de otra forma, cuando r^2 se aproxima a uno, los valores de y están cerca de la línea de regresión, por lo tanto, la variación total de los valores es más explicada por la línea, así la variable y está estrechamente relacionada con la variable x. Cuando la r^2 se aproxima a cero, los valores de y están lejos de la recta de regresión, por lo mismo, la variación total de los valores de y no es explicada en su mayor parte por la línea y la variable y no está estrechamente relacionada con la variable x.

Sin embargo, como r^2 siempre es un número positivo no da señal sobre si la relación entre 2 variables es positiva o negativa. Es por ello que se calcula la raíz cuadrada de r^2, $\pm r$ para mostrar el grado y la dirección de la relación. En el sentido de esta relación se toma del signo de la pendiente de la recta de regresión, b, que muestra si la relación entre las 2 variables es directa (+) o inversa (–).

Como el rango de r^2 va de 0 a 1, o $0 \le r^2 \le 1$, se tiene que el rango de r va de –1 a 1. En símbolos, $-1 \le r \le 1$, en donde el signo + indica una correlación positiva, en tanto que – indica una correlación negativa. La expresión que define a este coeficiente de correlación es:

$$r = \sqrt{\frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}} = \sqrt{\frac{SCR}{SCT}} \tag{13.22}$$

EJEMPLO 13.11

En el ejemplo sobre los ingresos y el ahorro, en la tabla 13.10 se describieron todas las sumas de cuadrados, se encontró que la suma de cuadrados de regresión SCR = 0.6020 y la suma de cuadrados total SCT = 0.916, por lo que el coeficiente de determinación es:

$$r^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = \frac{SCR}{SCT} = \frac{0.6020}{0.916} = 0.657$$

El coeficiente de correlación, como se desprende de la pendiente de la recta de regresión que se construyó para estos datos, tiene un signo positivo, ya que la recta sube de izquierda a derecha, y es:

$$r = \sqrt{0.657} = 0.81$$

Estos 2 valores muestran que la relación entre los datos y la recta de regresión es estrecha.

13.10.1 Coeficiente de correlación y Excel

Se ilustra aquí el uso del mecanismo "Coeficiente de correlación" de la sección de "Análisis de datos" de Excel.

Ejemplo 13.12

Existe un tema importante en finanzas, en particular en las inversiones en la Bolsa de Valores, se trata de la correlación entre el mercado y los rendimientos de las acciones que cotizan en la Bolsa. Con esta medida puede estimarse si el rendimiento de una acción sube (correlación positiva) o baja (correlación negativa) cuando el mercado en general sube, o viceversa. El mercado se mide a través de un índice de precios, como el Índice de Precios y Cotizaciones (IPC) de la Bolsa Mexicana de Valores (BMV).

Se evalúa aquí la correlación entre los rendimientos de este IPC con los de un par de acciones que cotizan en la BMV. El ejercicio es con pocas observaciones por cuestiones de espacio pero el lector interesado puede hacerlo con un mayo número, ya que es muy fácil obtener datos históricos en la sección de Finanzas de Yahoo! Los datos que utilizan aquí para el IPC provienen del sitio http://finance.yahoo.com/q/hp?s=%5EMXX+Historical+Prices, en tanto que los de C (Citibank) se tomaron de http://finance.yahoo.com/q/hp?a=&b=&c=&d=6&e=8&f=2011&g=d&s=c.mx. Mientras que los de Soriana B de http://finance.yahoo.com/q/hp?a=&b=&c=&d=6&e=8&f=2011&g=d&s=SORIANAB.MX. Todos fueron obtenidos el 8 de junio de 2011. En la tabla 13.13 se anotan los precios de cierre de esas 2 acciones y los valores del IPC.

Tabla 13.13 Valores del IPC de la BMV y precios de cierre de las acciones de C y Soriana.

Fecha	IPC	C	Soriana B
Mayo 9, 2011	35 467.01	512.80	37.25
Mayo 10, 2011	35 678.92	510.01	36.90
Mayo 11, 2011	35 380.53	501.00	37.50
Mayo 12, 2011	35 161.26	493.57	38.26
Mayo 13, 2011	35 045.14	487.10	37.14
Mayo 16, 2011	35 130.60	483.91	37.09
Mayo 17, 2011	34 819.57	486.15	37.90
Mayo 18, 2011	35 364.33	482.99	38.25
Mayo 19, 2011	35 276.47	481.00	37.00
Mayo 20, 2011	35 298.67	478.10	37.60
Mayo 23, 2011	35 215.02	470.89	37.05
Mayo 24, 2011	35 382.56	474.57	36.50
Mayo 25, 2011	35 498.42	471.54	36.02
Mayo 26, 2011	35 742.21	469.42	36.11
Mayo 27, 2011	35 819.20	475.65	36.50
Mayo 30, 2011	35 639.38	470.00	36.60
Mayo 31, 2011	35 832.79	475.20	36.00
Jun. 1, 2011	35 410.51	464.57	36.36
Jun. 2, 2011	35 416.25	466.26	35.60
Jun. 3, 2011	35 123.89	464.32	35.40
Jun. 6, 2011	34 673.48	447.40	34.66
Jun. 7, 2011	34 895.83	442.00	35.41
Jun. 8, 2011	34 879.07	436.00	34.90
Jun. 9, 2011	35 233.40	444.91	35.20
Jun. 10, 2011	34 963.78	451.64	34.88
Jun. 13, 2011	34 997.97	465.01	34.29
Jun. 14, 2011	35 445.65	457.34	35.10
Jun. 15, 2011	35 318.39	452.16	34.50
Jun. 16, 2011	35 220.99	449.87	34.20
Jun. 17, 2011	35 025.74	456.41	33.71
Jun. 20, 2011	35 109.97	452.80	33.50
Jun. 21, 2011	35 276.60	463.05	34.45
Jun. 22, 2011	35 399.44	466.40	34.12
Jun. 23, 2011	35 326.66	466.80	33.55
Jun. 24, 2011	35 347.85	471.70	33.76
Jun. 27, 2011	35 601.73	476.50	34.68
Jun. 28, 2011	36 188.91	474.50	35.15
Jun. 29, 2011	36 579.59	487.79	35.50
Jun. 30, 2011	36 558.07	487.49	35.11
Jul. 1, 2011	36 800.72	498.99	35.31
Jul. 4, 2011	36 847.10	500.00	35.25
Jul. 5, 2011	36 640.93	493.90	35.00
Jul. 6, 2011	36 468.01	487.57	34.85
Jul. 7, 2011	36 583.29	492.00	34.44

A partir de estos precios se calculan los rendimientos dividiendo el precio del día entre el precio del día anterior y restando 1 a este cociente, con lo que se obtienen los rendimientos diarios, en tanto por uno, que se muestran en la tabla 13.14.

Tabla 13.14 Precios y rendimientos de acciones y el IPC de la BMV.

Niveles / precios			Rendimientos diarios		
IPC	C	Soriana B	IPC	C	Soriana B
35 467.01	512.80	37.25			
35 678.92	510.01	36.90	0.0060	–0.0054	–0.0094
35 380.53	501.00	37.50	–0.0084	–0.0177	0.0163
35 161.26	493.57	38.26	–0.0062	–0.0148	0.0203
35 045.14	487.10	37.14	–0.0033	–0.0131	–0.0293
35 130.60	483.91	37.09	0.0024	–0.0065	–0.0013
34 819.57	486.15	37.90	–0.0089	0.0046	0.0218
35 364.33	482.99	38.25	0.0156	–0.0065	0.0092
35 276.47	481.00	37.00	–0.0025	–0.0041	–0.0327
35 298.67	478.10	37.60	0.0006	–0.0060	0.0162

(continúa)

Tabla 13.14 (*continuación*)

Niveles / precios			Rendimientos diarios		
IPC	C	Soriana B	IPC	C	Soriana B
35 215.02	470.89	37.05	–0.0024	–0.0151	–0.0146
35 382.56	474.57	36.50	0.0048	0.0078	–0.0148
35 498.42	471.54	36.02	0.0033	–0.0064	–0.0132
35 742.21	469.42	36.11	0.0069	–0.0045	0.0025
35 819.20	475.65	36.50	0.0022	0.0133	0.0108
35 639.38	470.00	36.60	–0.0050	–0.0119	0.0027
35 832.79	475.20	36.00	0.0054	0.0111	–0.0164
35 410.51	464.57	36.36	–0.0118	–0.0224	0.0100
35 416.25	466.26	35.60	0.0002	0.0036	–0.0209
35 123.89	464.32	35.40	–0.0083	–0.0042	–0.0056
34 673.48	447.40	34.66	–0.0128	–0.0364	–0.0209
34 895.83	442.00	35.41	0.0064	–0.0121	0.0216
34 879.07	436.00	34.90	–0.0005	–0.0136	–0.0144
35 233.40	444.91	35.20	0.0102	0.0204	0.0086
34 963.78	451.64	34.88	–0.0077	0.0151	–0.0091
34 997.97	465.01	34.29	0.0010	0.0296	–0.0169
35 445.65	457.34	35.10	0.0128	–0.0165	0.0236
35 318.39	452.16	34.50	–0.0036	–0.0113	–0.0171
35 220.99	449.87	34.20	–0.0028	–0.0051	–0.0087
35 025.74	456.41	33.71	–0.0055	0.0145	–0.0143
35 109.97	452.80	33.50	0.0024	–0.0079	–0.0062
35 276.60	463.05	34.45	0.0047	0.0226	0.0284
35 399.44	466.40	34.12	0.0035	0.0072	–0.0096
35 326.66	466.80	33.55	–0.0021	0.0009	–0.0167
35 347.85	471.70	33.76	0.0006	0.0105	0.0063
35 601.73	476.50	34.68	0.0072	0.0102	0.0273
36 188.91	474.50	35.15	0.0165	–0.0042	0.0136
36 579.59	487.79	35.50	0.0108	0.0280	0.0100
36 558.07	487.49	35.11	–0.0006	–0.0006	–0.0110
36 800.72	498.99	35.31	0.0066	0.0236	0.0057
36 847.10	500.00	35.25	0.0013	0.0020	–0.0017
36 640.93	493.90	35.00	–0.0056	–0.0122	–0.0071
36 468.01	487.57	34.85	–0.0047	–0.0128	–0.0043
36 583.29	492.00	34.44	0.0032	0.0091	–0.0118

Habiendo calculado los rendimientos, en la pestaña de "Datos" de la barra de herramientas de Excel puede activarse la sección "Análisis de datos" que aparece en el extremo derecho de esa barra, de ahí se desprende el menú con todas las alternativas, tal como se muestra en la figura 13.7.

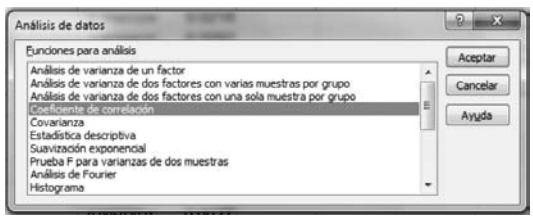


Figura 13.7 Menú del análisis de datos de Excel.

Al dar doble clic a "Coeficiente de correlación" aparece el siguiente cuadro de diálogo:

Si ahora se agrupan los datos bursátiles en 2 pares de columnas IPC – C e IPC – Soriana B, se puede correr el coeficiente de correlación 2 veces, una con cada par para obtener los resultados siguientes.

Para la correlación entre el IPC y las acciones de Citibank:

	Columna 1	Columna 2
Columna 1	1	
Columna 2	0.3897573	1

Para la correlación entre el IPC y las acciones de Soriana B:

	Columna 1	Columna 2
Columna 1	1	
Columna 2	0.30731565	1

Se observa que la correlación entre ambas acciones con el IPC es relativamente reducida, 0.39 para C y 0.31 para Soriana, sin embargo puede apreciarse que es mayor la correlación entre C y el IPC que entre éste y Soriana B.

13.10.2 Momento-producto de Pearson, otra manera de interpretar el coeficiente de correlación

El coeficiente de correlación puede abordarse como el coeficiente de correlación momento-producto de Pearson:

$$r = \frac{SC_{xy}}{\sqrt{SC_{xx}\,SC_{yyn}}} \tag{13.23}$$

Ejemplo 13.13

De vuelta al ejemplo 13.9, donde se construyó la ecuación de regresión de mínimos cuadrados para los datos de personas por hogar y los costos de bienes familiares, en la tabla 13.11 se resumieron los cálculos de las sumas de cuadrados y se encontró que SC_{xy} = 465 087.5, SC_{xx} = 82.5 y SC_{yy} = 2 674 103 174.10. Sustituyendo estos valores en la fórmula de Pearson:

$$r = \frac{SC_{xy}}{\sqrt{SC_{xx}\,SC_{yy}}} = \frac{465\,087.5}{\sqrt{82.5(2\,674\,103\,174.10)}}$$

$$= \frac{465\,087.5}{469\,695.13} = 0.99019$$

Se trata del mismo resultado que se obtiene mediante el "Coeficiente de correlación" del "Análisis de datos" de Excel:

	Columna 1	Columna 2
Columna 1	1	
Columna 2	0.99019018	1

Es también el mismo resultado que se obtiene a través de la fórmula (13.4), que se consiguió como la raíz cuadrada del cociente entre la suma de cuadrados de regresión y la suma de cuadrados total:

$$r = \sqrt{\frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}} = \sqrt{\frac{SCR}{SCT}}$$

En la tabla 13.12 se calcularon las sumas de cuadrados requeridas para encontrar el coeficiente de correlación por este camino:

$$r = \sqrt{\frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}} = \sqrt{\frac{SCR}{SCT}} = \sqrt{\frac{2\,621\,891\,601.15}{2\,674\,103\,174.10}}$$

$$= \sqrt{0.9804751} = 0.990189$$

En donde las pequeñas diferencias con los resultados obtenidos mediante la fórmula de momento-producto de Pearson y mediante Excel se deben a redondeos. Comparando las 2 maneras en las que puede calcularse el coeficiente de correlación:

$$r = \sqrt{\frac{\text{Variación explicada}}{\text{Variación total}}} = \sqrt{\frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}} = \sqrt{\frac{SCR}{SCT}}$$

$$r = \frac{SC_{xy}}{\sqrt{SC_{xx}\,SC_{YY}}}$$

Puede detectarse que la más sencilla es la del producto-momento de Pearson, ya que sólo requiere del cálculo de sumas de cuadrados a partir de los datos iniciales, en tanto que calcular ese coeficiente a partir del cociente de variaciones exige que, en primer lugar, se construya la ecuación de regresión para después determinar los valores estimados de y, $\hat{y}_i$. Luego a partir de esta ecuación se encuentran las sumas de cuadrados de regresión. Cabe mencionar que resulta mucho más claro comprender qué es lo que mide la correlación si se le contempla como cociente de variaciones.

13.10.3 Prueba de hipótesis sobre el coeficiente de correlación

Las hipótesis para esta prueba serían:

$$H_0 : \rho = 0$$

$$H_1 : \rho \neq 0$$

En donde ρ es la letra griega rho que representa el parámetro poblacional, el coeficiente de correlación de la población.

Aunque existe una prueba para evaluar estas hipótesis utilizando como estadístico de prueba la t de Student, para propósitos de sencillez basta con saber que esta prueba es equivalente a la revisada en la sección 13.8.1, donde se prueba si la pendiente de la recta es cero o no, ya que $\rho = 0$ sólo si la pendiente de la recta es igual a cero, es decir, si $\beta_1 = 0$.

Por ello, al evaluar si $\beta_1 = 0$ automáticamente también se está evaluando si $\rho = 0$, lo cual hace innecesario estudiar un mecanismo adicional de prueba.

13.10.4 Correlación serial o autocorrelación

En ocasiones se desea saber si las observaciones consecutivas de una serie de tiempo son aleatorias o si existe correlación entre observaciones cercanas. Esto puede evaluarse mediante la correlación serial, también conocida como *autocorrelación.*

Si se tiene una serie de tiempo, con observaciones $x_1, x_2 \ldots x_n$, la correlación serial se mide asociando a cada observación x_i el valor de x_{i+1}, de manera que se tienen los siguientes pares de valores:

x	y
x_1	x_2
x_2	x_3
…	…
x_j	x_{j+1}
…	…
x_{n-1}	x_n

El coeficiente de correlación que se calcula para estos pares de valores se denomina *coeficiente de correlación serial de orden 1*, ya que los valores que se asocian a los originales son los que le siguen.

De la misma manera pueden construirse coeficientes de correlación serial de orden 2, 3 o hasta k, según sea el adelanto de la observación que se asocia a los valores originales. Así, por ejemplo, se construiría una serie de pares de datos para calcular un coeficiente de correlación serial de orden 3 si al valor x_1 se le asocia x_4, al valor x_2 se le asocia el x_5, y así sucesivamente.

Este coeficiente de correlación serial o de autocorrelación sirve para probar si las observaciones de la serie de tiempo son aleatorias o no, funciona también para evaluar si existen características cíclicas en una serie de tiempo como, por ejemplo, si se desea evaluar si hay un efecto de ciclo de 6 años asociado con el cambio de presidente en México.

EJEMPLO 13.14

En la tabla 13.5 se muestran precios diarios de cierre para las acciones Alfa A, del 23 de mayo al 15 de julio de 2011; se han incluido en la tercera columna los datos de cierre con un día de retraso para calcular el coeficiente de correlación serial de orden 1.

Tabla 13.15 Datos de cierre de Alfa A, con una serie adicional con retraso 1.

Fecha	Precio de cierre	Precio de cierre con retraso de un periodo
23 mayo 2011	161.83	163.03
24 mayo 2011	163.03	163.15
25 mayo 2011	163.15	166
26 mayo 2011	166	167.98
27 mayo 2011	167.98	164
30 mayo 2011	164	167.1
31 mayo 2011	167.1	162.63
01 jun. 2011	162.63	162.11
02 jun. 2011	162.11	164.67
03 jun. 2011	164.67	163.5
06 jun. 2011	163.5	165.51
07 jun. 2011	165.51	166.4
08 jun. 2011	166.4	169.9
09 jun. 2011	169.9	169.7
10 jun. 2011	169.7	169.15
13 jun. 2011	169.15	175
14 jun. 2011	175	173
15 jun. 2011	173	174.3
16 jun. 2011	174.3	171.1
17 jun. 2011	171.1	172.5
06 jun. 2011	172.5	172.5
21 jun. 2011	172.5	171
22 jun. 2011	171	171.98
23 jun. 2011	171.98	171.25
24 jun. 2011	171.25	173
27 jun. 2011	173	174.35
28 jun. 2011	174.35	173.5
29 jun. 2011	173.5	174.5
30 jun. 2011	174.5	174
01 jul. 2011	174	173.55
04 jul. 2011	173.55	173
05 jul. 2011	173	174.75
06 jul. 2011	174.75	176.19
07 jul. 2011	176.19	178.4
08 jul. 2011	178.4	174.9
11 jul. 2011	174.9	173.8
12 jul. 2011	173.8	173.02
13 jul. 2011	173.02	172.25
14 jul. 2011	172.25	172.7
15 jul. 2011	172.7	

Nótese que originalmente había 40 datos pero al hacer la asociación con 1 periodo (día) de retraso ya sólo restan 39 pares.

Utilizando el mecanismo "Coeficiente de correlación" de Excel se obtiene un coeficiente de 0.89031463, un coeficiente de autocorrelación o de correlación serial, que muestra claramente que sí existe correlación entre cada dato y el que le sigue.

Ejemplo 13.15

Calcule el coeficiente de correlación serial para los datos de Alfa A, con un retraso de 2 y de 3 periodos.

Solución: En la tabla 13.16 se muestran los datos y las series con los retrasos solicitados.

Tabla 13.16 Datos de Alfa A, con series asociadas con retraso de 1 y 2 días.

Fecha	Precio de cierre	Precio de cierre con retraso de dos periodos	Precio de cierre con retraso de tres periodos
23 mayo 2011	161.83	163.15	166
24 mayo 2011	163.03	166	167.98
25 mayo 2011	163.15	167.98	164
26 mayo 2011	166	164	167.1
27 mayo 2011	167.98	167.1	162.63
30 mayo 2011	164	162.63	162.11
31 mayo 2011	167.1	162.11	164.67
01 jun. 2011	162.63	164.67	163.5
02 jun. 2011	162.11	163.5	165.51
03 jun. 2011	164.67	165.51	166.4
06 jun. 2011	163.5	166.4	169.9
07 jun. 2011	165.51	169.9	169.7
08 jun. 2011	166.4	169.7	169.15
09 jun. 2011	169.9	169.15	175
10 jun. 2011	169.7	175	173
13 jun. 2011	169.15	173	174.3
14 jun. 2011	175	174.3	171.1
15 jun. 2011	173	171.1	172.5
16 jun. 2011	174.3	172.5	172.5
17 jun. 2011	171.1	172.5	171
06 jun. 2011	172.5	171	171.98
21 jun. 2011	172.5	171.98	171.25
22 jun. 2011	171	171.25	173
23 jun. 2011	171.98	173	174.35
24 jun. 2011	171.25	174.35	173.5
27 jun. 2011	173	173.5	174.5
28 jun. 2011	174.35	174.5	174
29 jun. 2011	173.5	174	173.55
30 jun. 2011	174.5	173.55	173
01 jul. 2011	174	173	174.75
04 jul. 2011	173.55	174.75	176.19
05 jul. 2011	173	176.19	178.4
06 jul. 2011	174.75	178.4	174.9
07 jul. 2011	176.19	174.9	173.8
08 jul. 2011	178.4	173.8	173.02
11 jul. 2011	174.9	173.02	172.25
12 jul. 2011	173.8	172.25	172.7
13 jul. 2011	173.02	172.7	
14 jul. 2011	172.25		
15 jul. 2011	172.7		

Si se utiliza nuevamente el mecanismo "Coeficiente de correlación" de Excel se obtienen coeficientes de correlación serial de 0.85183685 y de 0.77678015 para los coeficientes de orden 2 y 3 respectivamente. Comparando los 3 coeficientes de correlación serial obtenidos en este ejemplo y en el anterior:

Orden (retraso)	Coeficiente de correlación serial
1	0.89031463
2	0.85183685
3	0.77678015

Puede verse que conforme aumenta el retraso (orden) el coeficiente de autocorrelación disminuye, era de esperarse, ya que resulta fácil visualizar que conforme más se alejan los datos entre sí menor es su relación, es decir, su correlación serial.

13.10.4.1 Prueba de hipótesis sobre el coeficiente de correlación serial

Tal como se señaló en la sección 13.10.3, una prueba sobre la significación estadística de este coeficiente de correlación serial puede llevarse a cabo a través de la prueba sobre si la pendiente de la recta de regresión es cero o no, $\beta_1 = 0$, revisada en la sección 13.7.1.

ejercicios 13.10 Análisis de correlación

1. Retome los 5 ejercicios de la sección 13.3 y siga las instrucciones que se muestran a continuación:

a) Calcule los coeficientes de determinación y de correlación.

i) Utilice la fórmula del cociente de variación explicada a variación total:

$$r^2 = \frac{\text{Variación explicada}}{\text{Variación total}} = \sqrt{\frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}} = \sqrt{\frac{SCR}{SCT}}$$

ii) Utilice la fórmula del producto-momento de Pearson:

$$r = \frac{SC_{xy}}{\sqrt{SC_{xx}\,SC_{yy}}}$$

b) Verifique con Excel que los resultados sean correctos.

2. Retome los ejercicios 6 a 10 y siga las instrucciones que se muestran a continuación:

a) Obtenga precios actuales de acciones que coticen en la Bolsa Mexicana de Valores (BMV) y calcule coeficientes de correlación serial de orden 1, 2 y 3.

b) Analice si el coeficiente de autocorrelación disminuye conforme aumenta el orden de la correlación. Los datos pueden obtenerse en el sitio web http://finance.yahoo.com/q/hp?a=&b=&c=&d=6&e=16&f=2011&g=d&s=amxl.mx que contiene los precios históricos de las acciones AMX L, de la empresa América Móvil.

En esta misma página web en la parte superior derecha aparece el recuadro siguiente:

En él se puede buscar la clave de cualquier acción que cotice en la Bolsa Mexicana de Valores (o de otras bolsas del mundo). Las siguientes son las claves de las acciones que más se negociaban para cuando se escribió este libro:

ALFAA	BOLSAA	FEMSAUBD	GFNORTEO	KIMBERA	TELMEXL
AMXL	C	GAPB	GMEXICOB	LABB	TLEVISACPO
ARA*	CEMEXCPO	GCARSOA1	GRUMAB	MEXCHEM	TVAZTCACPO
AXTELCPO	CHDRAUIB	GEOB	HOMEX	PENOLES	URBI
BIMBOA	ELEKTRA	GFINBURO	ICA*	SORIANAB	WALMEXV

Al anotar la clave de la acción deseada en el recuadro "Get historical prices for", deberá añadirle un punto y las letras mx, para que Yahoo! reconozca que se trata de acciones que se cotizan en la BMV.

13.11 Resumen

En este capítulo se estudiaron los temas del análisis de regresión y correlación lineal simples.

El análisis de regresión lineal simple se caracteriza porque se estudia la relación rectilínea (lineal) entre 2 variables (simple). Precisamente, el análisis comienza por determinar visualmente si puede considerarse que hay una relación rectilínea entre las 2 variables, se grafican los pares de valores en lo que se conoce como *diagrama de dispersión o nube de puntos*.

Si se determina la existencia de la relación lineal, entonces se procede a encontrar la ecuación que describirá a la recta de mejor ajuste a la nube de puntos. Aunque se revisó el *método de mano alzada* para dibujar la recta y encontrar la ecuación, también se detectó que el mejor método para hacerlo es el *método de mínimos cuadrados*, ya que permite asegurar que la suma de las distancias verticales entre cada uno de los puntos de datos y su correspondiente valor estimado, $\hat{y}_i$, es mínima.

Se revisaron 2 formas para derivar las 2 ecuaciones normales en las que se basa este método de mínimos cuadrados, algebraicamente y mediante derivadas parciales:

$$\sum y = na + b\sum x \qquad \textbf{(I)}$$

$$\sum xy = a\sum x + b\sum x^2 \qquad \textbf{(II)}$$

A su vez, se revisaron 4 métodos para resolver simultáneamente estas 2 ecuaciones y encontrar los parámetros de la ecuación de regresión:

1. Despejando simultáneamente a y b en las 2 ecuaciones normales.
2. Resolviendo simultáneamente las 2 ecuaciones normales.
3. Utilizando sumas de cuadrados.
4. Utilizando Excel.

Se aprendió que la forma más sencilla y conveniente para encontrar los parámetros de la ecuación de regresión es a través

de sumas de cuadrados, ya que además de permitir construir la ecuación, son datos (las sumas de cuadrados) que se utilizan para calcular otras cantidades útiles en el proceso de análisis: los errores estándar para las inferencias estadísticas, y para el cálculo de los coeficientes de correlación y de determinación.

Los mecanismos de Excel, como ya se ha visto, permiten ahorrar gran cantidad de esfuerzo en los abundantes cálculos que se requieren; sirven también para verificar que los cálculos realizados a mano (por ejemplo con calculadora) estén bien. Por supuesto se recomienda ampliamente realizar los cálculos a mano, ya que es la mejor manera de comprender lo que se está haciendo y el significado de los resultados.

Se revisaron además los supuestos en los que se basa el análisis de regresión:

1. x es una variable aleatoria.
2. La media de la distribución de probabilidad de ε, el término del error, es cero lo cual implica que el valor esperado del error es cero, es decir, $E(\varepsilon) = \Sigma(y_i - \hat{y}_i) = 0$.
3. La distribución de probabilidad de ε tiene forma normal.
4. Existe una subpoblación de valores y para cada valor de x; las subpoblaciones se distribuyen de forma normal.
5. Todas las varianzas de las subpoblaciones de y tienen la misma varianza.
6. Todas las medias de las subpoblaciones de y caen sobre la recta de regresión.

También se explicaron los procedimientos para realizar inferencias estadísticas: pruebas de hipótesis y estimaciones por intervalo de la pendiente de la recta de regresión, β_1; estimaciones y pronósticos por intervalo de valores de y para valores dados de x, que son de las aplicaciones más comunes del análisis de regresión.

Se aprendió que no se necesita estudiar un procedimiento adicional para evaluar si el coeficiente de correlación es igual a cero, ya que la prueba sobre si la pendiente de la recta es igual a cero es una prueba equivalente.

Todo lo anterior se estudió en secciones separadas. Para hacer un recuento y un repaso de corrido de lo que implica el análisis de regresión lineal, en la sección 13.10 se hizo una recapitulación del análisis de regresión lineal simple.

Finalmente, se revisó el tema del análisis de correlación mediante el cual se evalúa la intensidad y la dirección de la relación entre 2 variables. Se vio que el coeficiente de correlación puede calcularse de 2 maneras:

1. Como el cociente de variación explicada a variación total:

$$r = \sqrt{\frac{\text{Variación explicada}}{\text{Variación total}}} = \sqrt{\frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}} = \sqrt{\frac{SCR}{SCT}}$$

2. Con la fórmula del producto-momento de Pearson:

$$r = \frac{SC_{xy}}{\sqrt{SC_{xx}\,SC_{yy}}}$$

Se anotó que la más sencilla de calcular es la del producto-momento de Pearson, ya que sólo requiere del cálculo de sumas de cuadrados a partir de los datos iniciales, en tanto que, calcular ese coeficiente a partir del cociente de variaciones exige que se construya la ecuación de regresión para luego determinar los valores estimados de y, $\hat{y}_i$. Finalmente a partir de esta ecuación se encuentran las sumas de cuadrados de regresión.

Cabe mencionar que, a través de los cálculos por la fórmula del cociente de variaciones, resulta mucho más claro comprender qué es lo que mide la correlación si se le contempla como cociente de variaciones.

Finalmente, se estudió el coeficiente de correlación serial que permite evaluar si los datos de una serie de tiempo están relacionados entre sí, es decir, si están correlacionados.

13.12 Fórmulas del capítulo

13.1 Ecuación y recta de regresión

Procedimiento para encontrar la ecuación de una recta a partir de 2 puntos:

$$\frac{y_2 - y_1}{x_2 - x_1} = \frac{y - y_1}{x - x_1} \qquad \textbf{(13.1)}$$

Forma pendiente-ordenada al origen de una recta:

$$y = a + bx \qquad \textbf{(13.2)}$$

13.2 Método de mínimos cuadrados

Ecuaciones normales de mínimos cuadrados:

$$\sum y = na + b\sum x \qquad \textbf{(13.3)}$$

$$\sum xy = a\sum x + b\sum x^2 \qquad \textbf{(13.4)}$$

13.3 Determinación de la ecuación de regresión

Despeje simultáneo de a y b en las 2 ecuaciones normales:

$$a = \bar{y} - b\bar{x} \qquad \textbf{(13.5)}$$

$$b = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{\left(\sum x\right)^2}{n}} \qquad \textbf{(13.6)}$$

13.3.3 Resolución mediante sumas de cuadrados

Fórmulas para encontrar los parámetros de la ecuación de regresión mediante sumas de cuadrados:

$$a = \bar{y} - b\bar{x} \qquad \textbf{(13.5)}$$

$$b = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sum(x - \bar{x})^2} = \frac{SC_{xy}}{SC_{xx}} \qquad \textbf{(13.7)}$$

Sumas de cuadrados en el análisis de regresión:

13.5 Sumas de cuadrados en el análisis de regresión

$$\sum(y_i - \bar{y})^2 = \sum(y_i - \hat{y}_i)^2 + \sum(\hat{y}_i - \bar{y})^2 \qquad (13.8)$$

Suma de cuadrados total = suma de cuadrados del error + suma de cuadrados de regresión:

$$SCT = SCE + SCR \qquad (13.9)$$

13.6 Desviación estándar de regresión

Desviación estándar de regresión:

$$s = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{SCE}{n-2}} \qquad (13.10)$$

13.7 Inferencias estadísticas sobre la pendiente β_1

El error estándar, la desviación estándar de la distribución muestral de la pendiente de la recta de regresión, β_1:

$$\sigma_{\hat{\beta}_1} = \frac{\sigma}{\sqrt{SC_{xx}}} \qquad (13.11)$$

El estimador del error estándar, la desviación estándar de la distribución muestral de la pendiente de la recta de regresión, β_1:

$$s_{\hat{\beta}_1} = \frac{s}{\sqrt{SC_{XX}}} \qquad (13.12)$$

Estadístico de prueba t de Student para pruebas de hipótesis sobre la pendiente β_1:

$$t = \frac{\hat{\beta}_1 - \beta_1}{s\beta_1} \qquad (13.13)$$

Estadístico de prueba para la F de Fisher en pruebas de hipótesis sobre la pendiente β_1:

$$F = \frac{\frac{SCR}{1}}{\frac{SCE}{n-2}} = \frac{SCR(n-2)}{SCE} \qquad (13.14)$$

Una estimación por intervalo de β_1:

$$\hat{\beta} \pm ts_{\hat{\beta}_1} \qquad (13.15)$$

13.8.1 Estimación por intervalo de y para valores dados de x

Desviación estándar de la distribución muestral (error muestral) de $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$, para realizar estimaciones:

$$\sigma_{\hat{y}.e} = \sigma\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{SC_{xx}}} \qquad (13.16)$$

Intervalo de estimación de un valor de y para un valor dado de x:

$$\hat{y} \pm ts\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{SC_{xx}}} \qquad (13.17)$$

13.8.2 Pronósticos de y para valores dados de x

Desviación estándar de la distribución muestral (error muestral) de $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$, para realizar pronósticos:

$$\sigma_{\hat{y}.p} = \sigma\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{SC_{xx}}} \qquad (13.18)$$

Estimador de la desviación estándar de la distribución muestral (error muestral) de $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$, para realizar pronósticos:

$$s_{\hat{y}.p} = s\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{SC_{xx}}} \qquad (13.19)$$

El intervalo de pronóstico es:

$$\hat{y} \pm ts\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{SC_{xx}}} \qquad (13.20)$$

13.10 Análisis de correlación

El coeficiente de determinación como el cociente de 2 sumas de variaciones (cuadrados):

$$r^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = \frac{SCR}{SCT} \qquad (13.21)$$

El coeficiente de correlación como el cociente de 2 sumas de variaciones (cuadrados):

$$r = \sqrt{\frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}} = \sqrt{\frac{SCR}{SCT}} \qquad (13.22)$$

13.10.2 Momento-producto de Pearson, otra manera de interpretar el coeficiente de correlación

$$r = \frac{SC_{xy}}{\sqrt{SC_{xx}\,SC_{yy}}} \qquad (13.23)$$

13.13 Ejercicios adicionales

1. En la siguiente tabla se muestra el gasto en publicidad que realizaron 12 empresas durante un año y el volumen de ventas que obtuvieron.

Empresa	Gasto publicidad miles	Volumen de ventas miles
1	1 420	5 789
2	950	4 321
3	230	956
4	86	546
5	2 569	8 450
6	3 896	6 879
7	108	846
8	356	1 289

Empresa	Gasto publicidad miles	Volumen de ventas miles
9	789	3 021
10	956	4 258
11	456	1 876
12	520	2 200

a) Dibuje el diagrama de dispersión.
b) Trace a mano alzada la recta que mejor se ajuste a los puntos.
c) Determine la ecuación de regresión correspondiente a la recta trazada en *b*).
d) ¿Cuál es el volumen de ventas que se podría esperar para una inversión de $4 000 000 en publicidad?
e) Determine la ecuación de regresión con cualquiera de los 3 métodos de mínimos cuadrados que se presentaron.
f) Grafique esta recta de regresión de mínimos cuadrados en el diagrama de dispersión.
g) Verifique que se cumple el supuesto 2 del modelo: $E(\varepsilon) = \Sigma(y_i - \hat{y}_i) = 0$.
h) Realice la prueba de hipótesis sobre β_1 para determinar si existe relación entre las 2 variables; utilice un nivel de significación de 1 por ciento.
i) Haga una estimación por intervalo para β_1, con un nivel de confianza de 99 por ciento.
j) Estime, con un nivel de confianza de 99%, el volumen de ventas que se esperaría para una inversión en publicidad de $4 000 000.
k) Pronostique, con un nivel de confianza de 99%, el volumen de ventas que se esperaría para una inversión en publicidad de $4 000 000.
l) Calcule los coeficientes de determinación y de correlación.
m) Utilice la "Regresión" del "Análisis de datos" de Excel y verifique que los cálculos anteriores sean correctos.

2. Se quiere saber si existe una relación entre el ingreso familiar y el precio de la casa habitacional. Se tomó una muestra de 25 familias, los resultados se presentan en la siguiente tabla.

Familia	Ingreso (miles)	Precio casa (miles)
1	3.5	520
2	8.9	810
3	110	10 050
4	13	870
5	11.5	790
6	45	5 560
7	32	4 300
8	28	990
9	29.7	1 060
10	32.6	1 270
11	85	9 460
12	75.4	8 800
13	5.7	620
14	12	730
15	14.6	810
16	48.9	5 570
17	18.5	1 120
18	34.2	3 980
19	97	10 000
20	46	4 970
21	75.9	8 750
22	62.4	7 560
23	42	5 020
24	41	4 790
25	46	5 310

a) Dibuje el diagrama de dispersión.
b) Trace a mano alzada la recta que mejor se ajuste a los puntos.
c) Determine la ecuación de regresión correspondiente a la recta trazada en *b*).
d) ¿Cuál es el valor de la casa que se podría esperar para una familia que tiene ingresos de $100 000?
e) Determine la ecuación de regresión con cualquiera de los 3 métodos de mínimos cuadrados que se presentaron.
f) Grafique esta recta de regresión de mínimos cuadrados en el diagrama de dispersión.
g) Verifique que se cumple el supuesto 2 del modelo: $E(\varepsilon) = \Sigma(y_i - \hat{y}_i) = 0$.
h) Realice la prueba de hipótesis sobre β_1 para determinar si existe relación entre las 2 variables utilizando un nivel de significación de 0.05.
i) Haga una estimación por intervalo para β_1, con un nivel de confianza de 99 por ciento.
j) Estime, con un nivel de confianza de 99%, el valor de la casa para una familia que tiene ingresos de $100 000.
k) Pronostique, con un nivel de confianza de 99%, el valor de la casa para una familia que tiene ingresos de $100 000.
l) Calcule los coeficientes de determinación y de correlación.
m) Utilice la "Regresión" del "Análisis de datos" de Excel y verifique que los cálculos anteriores sean correctos.

3. En la siguiente tabla se muestran 9 pedidos hechos a una planta; las unidades producidas y su costo.

Pedido	Unidades producidas	Costo (miles)
1	400	53
2	320	41
3	250	37
4	106	28
5	700	102
6	453	58
7	565	76
8	620	83
9	200	24

a) Dibuje el diagrama de dispersión.
b) Trace a mano alzada la recta que mejor se ajuste a los puntos.
c) Determine la ecuación de regresión correspondiente a la recta trazada en *b*).
d) ¿Cuál es el costo de producción si se fabrican 1 000 unidades?

e) Determine la ecuación de regresión con cualquiera de los 3 métodos de mínimos cuadrados que se presentaron.
f) Grafique esta recta de regresión de mínimos cuadrados en el diagrama de dispersión.
g) Verifique que se cumple el supuesto 2 del modelo: $E(\varepsilon)=\Sigma(y_i-\hat{y}_i)=0$.
h) Realice la prueba de hipótesis sobre β_1 para determinar si existe relación entre las 2 variables, utilice un nivel de significación de 5 por ciento.
i) Haga una estimación por intervalo para β_1, con un nivel de confianza de 99 por ciento.
j) Estime, con un nivel de confianza de 99%, el costo de producción para una producción de 1 000 unidades.
k) Pronostique, con un nivel de confianza de 99%, el costo de producción para una producción de 1 000 unidades.
l) Calcule los coeficientes de determinación y de correlación.
m) Utilice la "Regresión" del "Análisis de datos" de Excel y verifique que los cálculos anteriores sean correctos.

4. En una cadena de comida rápida se desea saber si existe relación entre los ingresos obtenidos en un día y la población de comensales que laboran cerca de los restaurantes. Para analizarlo, se tomó una muestra de 10 restaurantes cercanos a oficinas y se determinó el número aproximado de clientes que pasan por sus establecimientos en cierto horario.

Restaurante	Población trabajadores	Ingresos diarios
1	400	52 200
2	348	47 150
3	285	36 520
4	104	24 100
5	47	12 300
6	523	67 890
7	149	30 250
8	96	28 790
9	265	31 850
10	302	42 500

a) Dibuje el diagrama de dispersión.
b) Trace a mano alzada la recta que mejor se ajuste a los puntos.
c) Determine la ecuación de regresión correspondiente a la recta trazada en *b*).
d) ¿Cuáles son los ingresos diarios esperados para una ubicación de restaurante en la que hay 600 comensales que laboran cerca?
e) Determine la ecuación de regresión con cualquiera de los 3 métodos de mínimos cuadrados que se presentaron.
f) Grafique esta recta de regresión de mínimos cuadrados en el diagrama de dispersión.
g) Verifique que se cumple el supuesto 2 del modelo: $E(\varepsilon)=\Sigma(y_i-\hat{y}_i)=0$.
h) Realice la prueba de hipótesis sobre β_1 para determinar si existe relación entre las 2 variables, con un nivel de significación de 5 por ciento.
i) Haga una estimación por intervalo para β_1, con un nivel de confianza de 99 por ciento.
j) Estime, con un nivel de confianza de 95%, los ingresos diarios esperados para una ubicación de restaurante en la que hay 600 comensales que laboran cerca.
k) Pronostique, con un nivel de confianza de 99%, los ingresos diarios esperados para una ubicación de restaurante en la que hay 600 comensales que laboran cerca.
l) Calcule los coeficientes de determinación y de correlación.
m) Utilice la "Regresión" del "Análisis de datos" de Excel y verifique que los cálculos anteriores sean correctos.

5. Se quiere saber si existe relación entre el salario de los trabajadores de una empresa y el ahorro que cada uno realiza. Se tomó una muestra de 10 trabajadores con los resultados que se muestran a continuación.

Trabajador	Sueldo	Ahorro
1	8 500	2 500
2	11 500	3 000
3	1 000	2 000
4	3 300	1 000
5	7 000	2 000
6	7 500	1 800
7	2 000	500
8	9 200	2 700
9	4 800	1 700
10	5 500	2 050

a) Dibuje el diagrama de dispersión.
b) Trace a mano alzada la recta que mejor se ajuste a los puntos.
c) Determine la ecuación de regresión correspondiente a la recta trazada en *b*).
d) ¿Cuál es el ahorro esperado para un trabajador con sueldo de $15 000?
e) Determine la ecuación de regresión con cualquiera de los 3 métodos de mínimos cuadrados que se presentaron.
f) Grafique esta recta de regresión de mínimos cuadrados en el diagrama de dispersión.
g) Verifique que se cumple el supuesto 2 del modelo: $E(\varepsilon)=\Sigma(y_i-\hat{y}_i)=0$.
h) Realice la prueba de hipótesis sobre β_1 para determinar si existe relación entre las 2 variables, utilice un nivel de significación de 5 por ciento.
i) Haga una estimación por intervalo para β_1, con un nivel de confianza de 99 por ciento.
j) Estime, con un nivel de confianza de 99%, el ahorro esperado para un trabajador con sueldo de $15 000.
k) Pronostique, con un nivel de confianza de 99%, el ahorro esperado para un trabajador con sueldo de $15 000.
l) Calcule los coeficientes de determinación y de correlación.
m) Utilice la "Regresión" del "Análisis de datos" de Excel y verifique que los cálculos anteriores sean correctos.

capítulo 14

Análisis de regresión lineal múltiple

Sumario

14.1 Modelo de regresión lineal múltiple y sus supuestos
14.2 Obtención de la ecuación de regresión lineal múltiple
14.3 Multicolinealidad y las variables que mejor se relacionan con la variable dependiente: uso de la matriz de correlaciones
14.4 Evaluación de la ecuación de regresión
14.4.1 Evaluación de la ecuación de regresión mediante el coeficiente de determinación múltiple
14.4.2 Evaluación de la ecuación de regresión mediante el análisis de varianza y la prueba *F*
14.4.3 Inferencias sobre coeficientes de regresión parciales individuales
14.4.4 Análisis de residuales
14.5 Uso del modelo de regresión lineal múltiple
14.5.1 Intervalos de confianza para los pronósticos
14.5.2 Intervalos de confianza para estimaciones de la media de una subpoblación de valores *y*
14.6 Variables independientes cualitativas
14.7 Regresión por pasos
14.7.1 Eliminación posterior
14.7.2 Regresión por pasos mediante selección previa
14.8 Resumen
14.9 Fórmulas del capítulo
14.10 Ejercicios adicionales

Así como en el análisis de regresión lineal simple se analiza la relación rectilínea entre 2 variables, una dependiente, y, y una independiente, x, en este análisis de regresión lineal múltiple se revisa la relación entre una variable y dependiente y 2 o más variables independientes $x_1, x_2, \ldots x_n$.

El modelo lineal simple que se estudió en el capítulo anterior:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Incluye:

β_0, la ordenada al origen,
β_1, la pendiente de la recta de regresión,
ε, el término del error, la diferencia entre cada uno de los valores observados de la variable y, y los valores estimados para cada valor de x, de acuerdo con la ecuación de regresión de mínimos cuadrados.

Por su parte, el modelo de regresión lineal múltiple se puede representar como:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_n x_n + \varepsilon \qquad (14.1)$$

en donde la diferencia con el modelo lineal simple consiste en los diferentes coeficientes β para cuantas (n) variables independientes se consideren en el modelo.

14.1 Modelo de regresión lineal múltiple y sus supuestos

El modelo de regresión lineal múltiple representado por

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_n x_n + \varepsilon$$

se basa en los supuestos siguientes:

1. Es un modelo lineal en los parámetros. En el capítulo 13 se mostró que el análisis de regresión lineal para 2 variables es lineal porque los exponentes de las variables de la ecuación que lo describe se

elevan a la potencia 1, lo cual, a su vez, indica que la gráfica de la ecuación es una línea recta. Sin embargo, por otro lado, tal como se verá en el capítulo 16, en la sección 16.3.3, una función como la exponencial, $y = f(x) = a^x$, se puede convertir en una ecuación lineal utilizando logaritmos, con lo que sus parámetros se vuelven lineales.

2. El modelo consta de una parte no aleatoria, la que se basa en las variables independientes, las x_i, y una parte aleatoria, los ε_i.
3. Para cada combinación de valores x_i existe una población de valores y_i que se distribuye en forma normal.
4. La varianza de todas la poblaciones de valores y_i que corresponden a cada combinación de valores x_i son todas iguales. A esta propiedad se le conoce como homoscedasticidad.
5. Los valores y_i son independientes entre sí. Esto significa que el valor observado para un valor de x no depende del valor observado para otro valor de x.
6. Los valores ε_i son independientes entre sí.
7. Los términos del error, las ε_i se distribuyen en forma normal con media de 0.

Homoscedasticidad. Propiedad que sostiene que para cada población de valor y_i le corresponderá una combinación de valor x_i.

Esta última suposición implica que el valor medio de y para un valor dado de x es:

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_n x_n$$

En secciones posteriores se harán pruebas para verificar que se cumplen algunos de estos supuestos.

14.2 Obtención de la ecuación de regresión lineal múltiple

El método que se aplica para obtener la ecuación de regresión lineal múltiple es el de mínimos cuadrados que se vio en el capítulo anterior, en cuya sección 13.2 se derivaron las ecuaciones normales, en las que se basa el método de mínimos cuadrados:

$$\sum y = na + b\sum x$$
$$\sum xy = a\sum x + b\sum x^2$$

Estas ecuaciones normales se utilizan para resolver aplicaciones de regresión lineal que implican sólo 2 variables: regresión lineal simple. Utilizando la simbología del modelo de regresión lineal, estas ecuaciones normales pueden escribirse como:

$$\sum y = nb_0 + b_1\sum x$$
$$\sum xy = b_0 \sum x + b_1 \sum x^2$$

Ya que, en la simbología del modelo de regresión lineal, b_0 es la pendiente de la recta y b_1 es el coeficiente de la variable independiente x. Es fácil ver la conveniencia de esta notación, ya que se puede extender la numeración de las b_i para incluir tantas variables independientes como sea necesario.

En la sección 13.2.1 del capítulo anterior se derivaron esas ecuaciones normales, sumando todos los términos de la ecuación lineal con 2 variables, con los que se obtuvo la primera ecuación normal y, después, se multiplicó esa misma ecuación lineal simple por x, la (única) variable independiente, se sumaron todos los términos, con lo que, finalmente, se obtuvo la segunda ecuación normal.

Siguiendo este mismo procedimiento se pueden obtener las ecuaciones normales para una ecuación lineal que tenga cualquier número de variables:

1. Se suman todos los términos de la ecuación lineal.
2. Se multiplica esa ecuación lineal por x_1, x_2, x_3, y así sucesivamente, según el número de variables de la ecuación lineal.
3. Se suman todos los términos de estas ecuaciones.

Se puede ilustrar el procedimiento para obtener las ecuaciones normales que se requiere resolver para abordar una aplicación de regresión lineal que tenga 2 variables independientes:

$$y = b_0 + b_1 x_1 + b_2 x_2$$

La suma de todos los términos de esta ecuación da:

$$\sum y = nb_0 + b_1 \sum x_1 + b_2 \sum x_2$$

que es la ecuación normal I.

Multiplicando ahora la ecuación lineal original por x_1 se obtiene:

$$x_1 y = b_0 x_1 + b_1 x_1^2 + b_2 x_1 x_2$$

Todos los términos de esta segunda ecuación sumados dan:

$$\sum x_1 y = b_0 \sum x_1 + b_1 \sum x_1^2 + b_2 \sum x_1 x_2 \ .$$

Con lo que se obtiene la segunda ecuación normal.

En tercer lugar, multiplicando la ecuación lineal con 2 variables independientes

$$y = b_0 + b_1 x_1 + b_2 x_2$$

por x_2, se obtiene

$$x_2 y = b_0 x_2 + b_1 x_1 x_2 + b_2 x_2^2$$

Finalmente, sumando todos los términos de esta ecuación, se obtiene

$$\sum x_2 y = b_0 \sum x_2 + b_1 \sum x_1 x_2 + b_2 \sum x_2^2$$

que es la tercera ecuación normal que se requiere para resolver una aplicación de regresión lineal múltiple con 2 variables independientes.

Así, el conjunto de ecuaciones normales que es necesario resolver simultáneamente en una aplicación de regresión lineal múltiple con 2 variables independientes es:

$$\sum y = nb_0 + b_1 \sum x_1 + b_2 \sum x_2 \tag{14.2}$$

$$\sum x_1 y = b_0 \sum x_1 + b_1 \sum x_1^2 + b_2 \sum x_1 x_2 \tag{14.3}$$

$$\sum x_2 y = b_0 \sum x_2 + b_1 \sum x_1 x_2 + b_2 \sum x_2^2 \tag{14.4}$$

El procedimiento para determinar la ecuación de regresión lineal múltiple en este caso (al igual que en el caso de la regresión simple) consiste en resolver simultáneamente estas 3 ecuaciones. Es relativamente sencillo hacer esto, aunque un tanto laborioso. Es claro que, conforme aumenta el número de variables independientes, la resolución se vuelve cada vez más complicada y, aunque existen métodos matriciales que ayudan con esta labor, dados los objetivos del presente texto, no se revisan aquí estas metodologías sino que se aprovechan las capacidades de Excel para resolver los ejemplos que aquí se presentan.

EJEMPLO 14.1

Los siguientes son algunos datos representativos de las nueve principales compañías de 2011, del listado de las 500 empresas más importantes de México, que cada año publica la revista *Expansión*.

#	Empresa	País	Ventas (mdp)	Activo	Pasivo	Patrimonio	Núm. de empleados
1	Pemex	MX	1 282 064.30	1 392 715.30	1 506 498.70	−113 783.40	147 672.00
2	América Móvil	MX	607 855.70	876 694.50	540 657.40	336 037.20	150 618.00
3	Walmart de México	EU	335 857.40	194 807.60	71 948.00	122 859.60	219 767.00
4	Comisión Federal de Electricidad	MX	254 417.30	841 202.30	488 545.50	352 656.80	93 254.00
5	Cemex	MX	178 260.00	515 097.00	301 397.00	213 700.00	46 523.00
6	Fomento Económico Mexicano	MX	169 701.80	223 578.40	70 565.30	153 013.10	108 572.00
7	General Motors de México	EU	158 692.00	55 191.00	42 073.00	13 112.00	12 000.00

(*continúa*)

(*continuación*)

#	Empresa	País	Ventas (mdp)	Activo	Pasivo	Patrimonio	Núm. de empleados
8	Grupo Alfa	MX	136 395.00	112 255.00	76 014.00	36 241.00	56 332.00
9	BBVA Bancomer	ESP	121 910.00	1 114 171.00	987 910.00	126 261.00	34 189.00

Fuente: CNN Expansión, *Las 500 empresas más importantes de México*, disponible en. http://www.cnnexpansion.com/rankings/2011/las-500-empresas-mas-importantes-de-mexico-2011/ranking.php, consultado el 7 de marzo de 2012.

Encuentre la ecuación de regresión utilizando las ventas como la variable dependiente y el activo, el pasivo, el patrimonio y el número de empleados como variables independientes.

Solución: El modelo aquí es:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

Se determinará la ecuación de regresión:

$$y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + b_4 x_4$$

Con base en los datos de la muestra de 9 empresas.

Utilizando el "Análisis de Datos" de Excel, con su sección de "Regresión", al igual que se hizo en el capítulo anterior, se obtienen los resultados que se muestran en la tabla 14.1.

Tabla 14.1 Resultados de Excel para el ejemplo 14.1

Estadísticas de la regresión						
Coeficiente de correlación múltiple			0.91521			
Coeficiente de determinación R^2			0.83761			
R^2 ajustado			0.67522			
Error típico			215 112			
Observaciones			9			
Análisis de varianza						
	Grados de libertad	Suma de cuadrados	Promedio de los cuadrados	*F*	Valor crítico de *F*	
Regresión	4	9.5472E+11	2.38681E+11	5.1581	0.07054536	
Residuos	4	1.8509E+11	46273153947			
Total	8	1.1398E+12				
	Coeficientes	Error típico	Estadístico *t*	Probabilidad	Inferior 95%	Superior 95%
Intercepción	–164 075.22	228 396.55	–0.72	0.51	–798 205.70	470 055.25
Activo	44 158.66	50 106.69	0.88	0.43	–94 959.82	183 277.14
Pasivo	–44 158.12	50 106.62	–0.88	0.43	–183 276.40	94 960.15
Patrimonio	–44 159.04	50 106.44	–0.88	0.43	–183 276.81	94 958.72
Empleo	3.17	1.30	2.45	0.07	–0.43	6.77

Se incluye toda la tabla (aunque, por otro lado, no se incluyeron todos los resultados que permite Excel) sólo para no perder de vista toda la información que este mecanismo del paquete de Microsoft produce.

Sin embargo, por el momento, sólo interesan los coeficientes de la ecuación de regresión lineal múltiple.

Así, con:

x_1 = activo,
x_2 = pasivo,
x_3 = patrimonio,
x_4 = empleados

la ecuación de regresión lineal múltiple es:

$$y = -164\,075.22 + 44\,158.66x_1 - 44\,158.12x_2 - 44\,159.04x_3 + 3.17x_4$$

En esta ecuación de regresión, los coeficientes de regresión estimados, los coeficientes de todas las variables x_i, miden el cambio promedio en la variable dependiente debido a un incremento de una unidad en la variable predictora correspondiente, manteniendo constantes todas las otras variables de predicción.

Es importante resaltar un punto que se señala en el párrafo anterior y que se refiere a la interpretación de los coeficientes de regresión: en la regresión simple, existe una sola variable independiente, x, en tanto que, en una regresión lineal múltiple existen cuando menos 2, que serían x_1 y x_2, o 4, como en el ejemplo anterior. En la regresión simple se interpreta a la variable x como el cambio que se da en la variable dependiente, y, por los cambios que se dan en esa variable independiente x. Por su parte, en regresión múltiple, la interpretación de cualquier variable independiente trata de la variación que se da en y, por la variación que se da en esa variable independiente específica, si se mantienen todas las demás sin cambio. Por ejemplo, en la ecuación de regresión anterior, la x_1 representa los cambios que el activo ocasiona en la variable dependiente, las ventas, si las otras 3 variables independientes se conservan sin cambio.

Antes de poder utilizar la ecuación de regresión múltiple para sus principales usos, estimación y pronósticos de valores de la variable dependiente, es necesario asegurarse de que se tiene un modelo adecuado. Por ello, en las secciones siguientes se revisan diversos procedimientos para evaluar si esta ecuación de regresión es adecuada en términos de diversos criterios.

14.3 Multicolinealidad y las variables que mejor se relacionan con la variable dependiente: uso de la matriz de correlaciones

La **multicolinealidad**, un problema que se debe evitar en los análisis de regresión, se da cuando las variables independientes están altamente correlacionadas. Esta multicolinealidad puede ocasionar problemas en el uso de la ecuación de regresión para hacer estimaciones o pronósticos sobre la variable dependiente y, como es prácticamente armar un conjunto de variables independientes que no se correlacionan, por lo general se considera que un par de variables independientes son aceptables si su correlación no es mayor de 0.7, en sentido positivo o en sentido negativo. Para evaluar esto, conviene calcular al inicio del análisis un matriz de correlaciones para ver el índice de correlación que existe entre cada par de variables independientes.

Multicolinealidad. Es un problema que se da cuando las variables independientes están altamente correlacionadas.

Adicionalmente, la matriz de correlaciones sirve para elegir las variables que mejor se relacionan con la variable dependiente ya que, como se recordará, el coeficiente de correlación mide qué tan estrecha es la relación entre 2 variables. En el ejemplo siguiente se analizan estos aspectos.

EJEMPLO 14.2

Utilizando los datos de las 9 empresas mexicanas más importantes de 2011, calcule la matriz de correlación para revisar la relación entre las ventas como variable dependiente y las restantes como variables independientes.

Solución: En la tabla 14.2 se muestran los resultados obtenidos utilizando el mecanismo de "Coeficiente de correlación" de la sección de "Análisis de datos" de Excel.

Tabla 14.2 Matriz de correlación para los datos de las 9 empresas mexicanas más importantes de 2011

	Ventas (mdp)	Activo	Pasivo	Patrimonio	Núm. de empleados
Ventas (mdp)	1				
Activo	0.661695352	1			
Pasivo	0.748428863	0.95441734	1		
Patrimonio	–0.377170258	0.02172925	–0.27766612	1	
Núm. de empleados	0.520962336	0.14659829	0.1148404	0.08718297	1

En primer lugar, note que sólo tiene datos la mitad izquierda inferior de la tabla, ya que la otra mitad corresponde a los mismos pares de variables; además, la diagonal siempre equivaldrá a 1, ya que la correlación de cualquier variable consigo misma es perfecta.

Pasando ahora a los coeficientes de correlación calculados para cada par de variables, se observa que el mayor es 0.95441734 que corresponde a la correlación entre el activo y el pasivo, y le siguen, en orden de mayor a menor, los coeficientes de correlación para ventas y pasivo, ventas y activo y, en cuarto lugar, ventas y número de empleados.

Analizando lo anterior, se decidiría eliminar del análisis la variable del activo ya que, al estar estrechamente correlacionada con el pasivo, puede provocar problemas de multicolinealidad

y también se puede asumir que ambas aportan información similar a las ventas. Siendo el activo la variable que está menos correlacionada con las ventas, se le elimina a favor del pasivo y, como éste es el único par de variables independientes que tiene una correlación superior a 0.70, se pasa a la segunda parte del análisis de la matriz de correlaciones.

Ahora se elige eliminar también la variable de patrimonio, ya que es la que tiene la menor correlación con las ventas.

EJEMPLO 14.3

Tras hacer lo anterior se procede a calcular una nueva ecuación de regresión para los datos de las 9 empresas, con lo que los datos se reducen a:

#	Empresa	País	Ventas (mdp)	Pasivo	Núm. de empleados
1	Pemex	MX	1 282 064.30	1 506 498.70	147 672.00
2	América Móvil	MX	607 855.70	540 657.40	150 618.00
3	Walmart de México	EU	335 857.40	71 948.00	219 767.00
4	Comisión Federal de Electricidad	MX	254 417.30	488 545.50	93 254.00
5	Cemex	MX	178 260.00	301 397.00	46 523.00
6	Fomento Económico Mexicano	MX	169 701.80	70 565.30	108 572.00
7	General Motors de México	EU	158 692.00	42 073.00	12 000.00
8	Grupo Alfa	MX	136 395.00	76 014.00	56 332.00
9	BBVA Bancomer	ESP	121 910.00	987 910.00	34 189.00

Fuente: CNN Expansión, *Las 500 empresas más importantes de México*, disponible en: http://www.cnnexpansion.com/rankings/2011/las-500-empresas-mas-importantes-de-mexico-2011/ranking.php/, consultado el 12 de marzo de 2011.

Utilizando de nuevo "Regresión" del "Análisis de datos" de la opción de "Datos" de Excel, se obtienen los siguientes coeficientes (ya sólo se anotan éstos y no toda la tabla):

	Coeficientes
Intercepción	–116 147.541
Pasivo	0.522836491
Núm. de empleados	2.47935207

Por lo que la nueva ecuación de regresión lineal múltiple es:

$$y = -166\,147.541 + 0.52283649x_1 + 2.47935207x_2$$

Resumiendo lo anterior, se puede concluir que:

1. Ningún par de variables independientes debe tener una correlación inferior a –0.70 ni superior a 0.70.
2. Una variable independiente, a las que también se les denominan **predictoras**, cuando se les utiliza para hacer pronósticos sobre la variable dependiente, debe tener, preferentemente, una correlación fuerte con la variable dependiente.

Variable predictora. Es una variable independiente que se utiliza para hacer pronósticos sobre la variable dependiente.

EJERCICIOS 14.3 Multicolinealidad y las variables que mejor se relacionan con la variable dependiente: uso de la matriz de correlaciones

1. Analice al siguiente matriz de correlaciones, en donde la variable A es la variable dependiente.

	A	B	C	D	E
A	1				
B	0.912	1			
C	0.258	0.187	1		
D	0.423	0.397	–0.110	1	
E	0.525	0.095	–0.565	0.815	1

a) ¿Cuál variable incorporaría usted primero al modelo? ¿Por qué?

b) ¿Cuál variable incorporaría usted en segundo lugar al modelo? ¿Por qué?

c) ¿Cuál o cuáles variables no incorporaría usted al modelo? ¿Por qué?

2. Analice la siguiente matriz de correlaciones, en donde la variable A es la variable dependiente.

	A	B	C	D
A	1	–0.865	0.794	0.225
B	–0.865	1	–0.416	–0.082
C	0.794	–0.416	1	0.11
D	0.225	–0.082	0.11	1

a) ¿Cuál variable incorporaría usted primero al modelo? ¿Por qué?
b) ¿Cuál variable incorporaría usted en segundo lugar al modelo? ¿Por qué?
c) ¿Cuál o cuáles variables no incorporaría usted al modelo? ¿Por qué?

3. En la tabla siguiente se encuentran datos sobre PIB y distribución de las poblaciones en las actividades económicas primaria, secundaria y terciaria para 45 países incluyendo México, obtenidos de www.portalplanetasedna.com.ar/datos03.htm/, consultado el 7 de marzo de 2012.

País	PIB (USD)	% Población urbana	% Act. Econ. Prim.	% Act. Econ. Sec.	% Act. Econ. Terc.
Antigua y Barbuda	738	73.5	4.0	18.9	77.1
Argentina	12 277	73.2	5.7	28.7	65.6
Barbados	14353	76.6	6.6	20.0	73.4
Belice	4 959	73.8	18.7	25.5	55.8
Bolivia	2 355	62.0	15.4	28.7	55.9
Brasil	7037	67.5	8.4	28.8	62.8
Canadá	26 251	78.7	3.0	31.0	66.0
Chile	8 652	75.2	7.4	30.4	62.2
Colombia	5 749	70.9	13.5	25.1	61.4
Costa Rica	8 860	76.2	15.2	24.3	60.5
Cuba	1170	75.5	7.0	31.0	62.0
Dominica	3040	77.5	20.2	22.5	57.3
Ecuador	2 994	64.3	12.9	35.2	51.9
El Salvador	4 344	46.3	12.1	28.0	59.9
EE.UU.	31 872	77.0	1.7	26.2	72.0
Granada	314	36.6	8.4	22.2	69.4
Guatemala	3 674	39.4	23.3	20.0	56.8
Guyana	364	37.6	34.7	32.5	32.8
Haití	1 464	35.1	30.4	20.1	49.6
Honduras	234	51.6	20.3	30.9	48.8
Jamaica	3 561	55.6	8.0	33.7	58.4
México	8 297	74.2	4.9	26.6	68.4
Nicaragua	2 279	55.8	34.1	21.5	44.4
Panamá	5 875	56.0	7.9	18.4	73.8
Paraguay	4 384	55.3	24.9	26.2	48.9
Perú	4 622	72.4	7.1	36.8	56.1
Puerto Rico	200	67.1	1.0	44.0	55.0
Rep. Dominicana	5 507	64.4	11.6	32.8	55.6
Saint Kitts y Nevis	626	42.4	4.6	24.3	71.1
San Vicente y Granadinas	242	50.9	10.9	26.9	62.2
Santa Lucía	351	48.1	8.1	18.9	72.9
Somalia	110	37.2	64.0	8.0	28.0
Sudáfrica	8 908	50.1	4.0	31.8	64.3
Sudán	664	35.1	39.3	18.2	42.6
Surinam	4 178	73.5	12.0	29.0	59.0
Swazilandia	3 987	26.1	16.0	38.7	45.3
Tanzanía	501	31.6	14.9	39.4	45.7
Togo	1410	51.6	42.1	21.1	36.8
Trinidad y Tobago	8 176	73.6	1.8	47.5	50.7
Túnez	5957	69.9	12.4	28.4	59.1
Uganda	1167	43.2	44.6	17.6	37.8
Uruguay	8 879	91.0	8.5	27.5	64.0
Venezuela	5 495	86.6	5.0	34.0	61.0
Zambia	756	41.0	17.3	26.4	56.3
Zimbabwe	2876	42.9	19.5	24.4	56.1

Calcule la matriz de correlaciones:

a) ¿Cuál variable incorporaría usted primero al modelo? ¿Por qué?
b) ¿Cuál variable incorporaría usted en segundo lugar al modelo? ¿Por qué?
c) ¿Cuál o cuáles variables no incorporaría usted al modelo? ¿Por qué?
d) Determine la ecuación de regresión para el mejor modelo posible, con el PIB como variable dependiente.

4. Una empresa de paquetería evalúa los tiempos de entrega, por lo que recaba información sobre los tiempos de recorrido de rutas junto con las distancias recorridas y el número de paquetes entregados. En la tabla siguiente se muestran los datos.

Ruta	Tiempo del recorrido (horas)	Recorrido (km)	Paquetes entregados
1	10.4	165	5
2	5.7	82	4
3	9.7	165	5
4	7.6	165	3
5	5.2	82	3
6	7.3	132	2
7	8.3	124	4
8	7	107	4
9	8.5	149	4
10	7.1	149	3
11	6.7	132	4
12	6.1	140	2

a) Construya la matriz de correlaciones y comente.
b) Determine la ecuación de regresión para el mejor modelo posible, con el tiempo de recorrido como variable dependiente.

5. En la tabla siguiente se presenta una lista con diversas características de 6 autos híbridos que se venden en Estados Unidos.

Marca	Modelo	Precio (USD)	Millas por galón en ciudad	Millas por galón en carretera	Potencia (hp)	rpm
Toyota	Prius	20 875	61	50	76	5 000
Honda	Civic	20 650	48	47	93	5 700
Honda	Accord	29 990	30	37	255	6 000
Ford	Escape	26 780	36	31	133	6 000
Lexus	RX 400h	49 060	31	27	268	5 600
Toyota	Highlander	39 855	33	28	268	5 600

Fuente: http://www.allabouthybridcars.com/comparison-chart.htm/, 7 de marzo de 2012.

a) Calcule la matriz de correlaciones.
b) ¿Cuál variable incorporaría usted primero al modelo? ¿Por qué?
c) ¿Cuál o cuáles variables no incorporaría usted al modelo? ¿Por qué?
d) Determine la ecuación de regresión para el mejor modelo posible, utilizando el precio como variable dependiente.

14.4 Evaluación de la ecuación de regresión

Antes de poder utilizar la ecuación de regresión en pronósticos o estimaciones sobre la variable dependiente, se le debe evaluar en 3 sentidos. En primer lugar, se le evalúa en forma global para medir qué tan adecuada es para medir la relación entre las variables. Esta evaluación se hace mediante:

1. El coeficiente de determinación múltiple.
2. El análisis de varianza y la prueba con el estadístico F de Fisher.

Un segundo tipo de prueba sobre la ecuación de regresión es la que se hace sobre los coeficientes de regresión parciales individuales, con el propósito de evaluar si son estadísticamente significativos o, en otras palabras, para ver si son diferentes de 0 porque, en el caso de no poder probar que lo son, entonces las variables asociadas no son útiles en el modelo.

El tercer tipo de evaluación que se hace de la ecuación de regresión es la que se realiza sobre los residuales, la diferencia entre los valores observados de la variable dependiente y los valores estimados a través de la ecuación de regresión lineal múltiple que se construye. Esta evaluación de los residuales se relaciona, más que con la adecuación en sí del modelo, con el cumplimiento de los supuestos en los que se basa.

En los apartados siguientes se analizan los 3 tipos de constataciones.

14.4.1 Evaluación de la ecuación de regresión mediante el coeficiente de determinación múltiple

Este coeficiente de determinación múltiple, r^2, proporciona una medida global de qué tan adecuada es la ecuación para medir la relación entre las variables. Se vio en la sección 13.10 del capítulo anterior que el coeficiente de determinación se define como la razón de la variación explicada a la variación total:

$$r^2 = \frac{Variación\ explicada}{Variación\ total} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = \frac{SCR}{SCT}$$

Este coeficiente de determinación múltiple indica qué proporción de la variabilidad total de la variable dependiente sería explicada por las variables independientes.

Sin embargo, en este caso, como r^2 es un estimador positivamente sesgado del correspondiente parámetro poblacional, se le ajusta de la siguiente manera:

$$r^2_{aj} = 1 - \left[\left(1 - r^2\right)\frac{n-1}{n-k-1}\right] \qquad (14.5)$$

Ejemplo 14.4

Para el ejemplo del efecto que tienen 4 variables sobre las ventas de las 9 empresas más importantes de México, en la tabla 14.1 se muestran los datos de las sumas de cuadrados que se requieren para calcular el coeficiente de determinación r^2:

$$r^2 = \frac{\textit{Variación explicada}}{\textit{Variación total}} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = \frac{SCR}{SCT}$$

$$= \frac{954\ 720\ 000\ 000}{1\ 139\ 800\ 000\ 000} = 0.84$$

Esto señala que el modelo es bueno, ya que 84% de la variación total queda explicada por la regresión.

Ejemplo 14.5

Retomando el modelo reducido del ejemplo 14.3 en el que se corrió la regresión para las ventas de las 9 empresas mexicanas más importantes, utilizando como variables independientes sólo el pasivo y el número de empleados, se llegó a la ecuación siguiente:

$$y = -116\ 147.541 + 0.52283649x_1 + 2.47935207x_2$$

Al utilizar de nuevo el mecanismo "Regresión" del "Análisis de datos" de Excel se obtienen los siguientes datos de las sumas de cuadrados:

Análisis de varianza		
	Grados de libertad	Suma de cuadrados
Regresión	2	8.5704E+11
Residuos	6	2.8278E+11
Total	8	1.1398E+12

De donde, el coeficiente de determinación es:

$$r^2 = \frac{\textit{Variación explicada}}{\textit{Variación total}} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = \frac{SCR}{SCT}$$

$$= \frac{857\ 040\ 000\ 000}{1\ 139\ 800\ 000\ 000} = 0.752$$

Nótese la notación científica de las sumas de cuadrados de la tabla anterior: 8.5704E+11, equivale a 857 040 000 000 en notación decimal; es decir, el "+11" de la notación científica indica que se debe correr el punto decimal de "8.5704" once posiciones a la derecha.

Ese r^2 de 0.752 señala un modelo igualmente aceptable como el que utiliza las 4 variables independientes, pero con una menor proporción de variación explicada aunque, por otro lado, con sólo 2 variables independientes. En otras palabras, el modelo con 4 variables independientes explica 84% de la variación total (ejemplo 14.3), en tanto que este modelo con sólo 2 variables explica 75.2%, lo cual quiere decir que las 2 variables eliminadas dan cuenta del 84 – 75.2 = 8.8% de la variación.

14.4.2 Evaluación de la ecuación de regresión mediante el análisis de varianza y la prueba *F*

El análisis de varianza y la prueba *F* se utilizan también para realizar una prueba global de significación para la regresión. En primer lugar se plantean las hipótesis:

- $H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$, no existe relación lineal entre y y el conjunto de variables independientes.
- H_1: Por lo menos una de las betas es diferente de o: sí existe relación lineal entre y y cuando menos una de las variables independientes.

El estadístico de prueba:

$$F = \frac{\textit{Cuadrados medios de regresión}}{\textit{Cuadrados medios de error}} = \frac{CMR}{CME}$$

Ejemplo 14.6

Volviendo al ejemplo de regresión sobre las ventas de las 9 empresas más importantes de México utilizando: *a*) 4 variables independientes, *b*) 2 variables independientes. Los datos más relevantes de los resultados que se obtienen con Excel:

a)

Análisis de varianza					
	Grados de libertad	Suma de cuadrados	Promedio de los cuadrados	*F*	Valor crítico de *F*
Regresión	4	954 720 000 000	238 681 000 000	5.1581	0.07054536
Residuos	4	185 090 000 000	46 273 153 947		
Total	8	1 139 800 000 000			

El valor empírico del estadístico de prueba:

$$F = \frac{\textit{Cuadrados medios de regresión}}{\textit{Cuadrados medios de error}} = \frac{CMR}{CME}$$

$$= \frac{238\ 681\ 000\ 000}{46\ 273\ 153\ 947} = 5.1581$$

Que es, por cierto, otro valor que Excel produce, junto con un "Valor crítico de *F*", el cual representa la probabilidad de obtener ese valor de 5.1581 para el estadístico de prueba; a su vez, como esta probabilidad es mayor que el nivel de significación de 0.01, no se rechaza la hipótesis nula y se concluye que no hay relación lineal entre *y* y el conjunto de variables independientes.

b)

	Grados de libertad	Suma de cuadrados	Promedio de los cuadrados	*F*	Valor crítico de *F*
Regresión	2	857 040 131 329	428 520 065 664	9.09241832	0.01526948
Residuos	6	282 776 298 263	47 129 383 044		
Total	8	1 139 816 429 592			

Como puede verse en la tabla anterior, el valor empírico de la *F* es 9.092 y, como la probabilidad de obtener este valor es de 10.925, que también es superior al nivel de significación de 1%, no es posible rechazar la hipótesis nula y se concluye, al igual que con el modelo con 4 variables independientes, que no existe relación lineal entre *y* el conjunto de 2 variables independientes.

A la luz de estos resultados, no se aborda ya el caso de las 9 empresas más importantes de México ya que, con los datos disponibles, no se puede obtener un buen modelo de regresión lineal múltiple.

Para resumir lo visto hasta aquí y para continuar revisando los temas relacionados con el análisis de regresión lineal múltiple, se utiliza otro ejemplo.

EJEMPLO 14.7

Los datos de la tabla 14.3 corresponden a una muestra de 6 000 hogares con ingresos inferiores a $180 000 anuales. Los hogares se clasificaron en 36 grupos demográficos con el propósito de evaluar los salarios diarios. Haga un análisis de regresión.

Tabla 14.3 Datos para el ejemplo 14.7

Grupo demográfico	Promedio de activos bancarios de la familia	Horas trabajadas	Salario diario	Ingreso anual promedio del cónyuge	Ingreso anual promedio de otros miembros de la familia	Edad del jefe de familia	Promedio del número de dependientes económicos	Promedio del grado máximo de escolaridad
1	23 325.00	1 985	142.30	6 912.50	4 762.50	40.60	3.833	6.6
2	21 737.50	2 042	230.40	13 562.50	4 100.00	41.80	2.444	8.2
3	24 837.50	2 047	245.30	15 162.50	3 712.50	40.30	2.545	9.1
4	35 075.00	2 051	257.30	14 925.00	3 487.50	40.00	2.362	9.1
5	38 350.00	2 062	235.00	15 175.00	4 075.00	40.10	2.851	8.9
6	55 487.50	2 067	290.90	22 562.50	3 625.00	39.10	2.301	10.5
7	17 125.00	2 077	190.10	4 375.00	2 612.50	37.40	4.158	8.2
8	70 662.50	2 084	298.30	16 587.50	4 137.50	39.80	2.208	10.2
9	17 812.50	2 093	189.90	4 275.00	3 887.50	37.50	4.512	8.1
10	55 000.00	2 098	228.00	12 162.50	4 550.00	40.60	2.661	8.4

Grupo demográfico	Promedio de activos bancarios de la familia	Horas trabajadas	Salario diario	Ingreso anual promedio del cónyuge	Ingreso anual promedio de otros miembros de la familia	Edad del jefe de familia	Promedio del número de dependientes económicos	Promedio del grado máximo de escolaridad
11	94 462.50	2 102	323.40	14 850.00	5 175.00	39.80	2.019	10.7
12	59 125.00	2 105	249.30	14 750.00	3 875.00	39.90	2.616	9.3
13	57 675.00	2 108	279.60	12 950.00	3 750.00	38.20	2.04	9.2
14	63 362.50	2 109	249.90	15 087.50	4 337.50	39.70	3.193	8.9
15	20 400.00	2 111	251.10	15 037.50	612.50	22.40	1.159	11.5
16	73 837.50	2 121	292.20	15 637.50	4 100.00	39.80	2.287	10.3
17	100 525.00	2 127	326.20	15 325.00	3 925.00	39.50	2.259	10.8
18	158 875.00	2 134	279.10	12 662.50	7 425.00	57.70	1.229	8.8
19	90 625.00	2 157	290.50	14 012.50	3 637.50	38.50	2.34	10.5
20	70 262.50	2 159	251.10	13 437.50	3 612.50	39.30	2.486	9.5
21	95 312.50	2 173	295.90	13 950.00	3 700.00	39.20	2.342	10.5
22	96 800.00	2 174	297.00	14 100.00	3 762.50	39.30	2.335	10.5
23	127 987.50	2 174	358.20	14 262.50	5 175.00	40.00	2.064	11.7
24	97 237.50	2 179	297.10	14 100.00	3 900.00	39.40	2.341	10.5
25	91 750.00	2 181	291.20	13 400.00	3 800.00	39.00	2.337	10.2
26	140 500.00	2 184	363.60	13 637.50	3 637.50	39.10	2.328	11.6
27	96 325.00	2 185	304.00	14 187.50	3 587.50	38.60	2.602	10.7
28	91 150.00	2 186	301.50	14 025.00	375.00	37.20	2.046	10.9
29	91 562.50	2 188	301.00	12 375.00	4 575.00	38.40	2.847	10.6
30	86 100.00	2 196	300.90	11 837.50	3 675.00	37.50	3.047	10.6
31	130 625.00	2 197	341.30	13 475.00	3 750.00	39.10	2.297	11.3
32	98 562.50	2 200	298.00	14 075.00	2 550.00	39.20	2.341	10.6
33	84 862.50	2 205	235.60	11 062.50	3 300.00	38.80	2.662	9.5
34	116 725.00	2 210	322.20	13 750.00	3 687.50	39.00	2.187	11.2
35	91 162.50	2 257	251.60	13 662.50	2 200.00	37.90	2.042	10.1
36	103 962.50	2 267	283.80	16 225.00	3 150.00	38.90	2.024	11.1

Solución: En la tabla 14.4 se muestran los resultados que se obtuvieron con la "Regresión del análisis de datos" de Excel.

Tabla 14.4 Resultados de Excel para el ejemplo 14.7

Estadísticas de la regresión						
Coeficiente de correlación múltiple			0.96859807			
Coeficiente de determinación R^2			0.93818222			
R^2 ajustado			0.92272778			
Error típico			10 145.5817			
Observaciones			36			
Análisis de varianza						
	Grados de libertad	Suma de cuadrados	Promedio de los cuadrados	*F*	Valor crítico de *F*	
Regresión	7	4.3741E+10	6 248 671 538	60.706304	3.0321E-15	
Residuos	28	2882119179	102 932 828			
Total	35	4.6623E+10				
	Coeficientes	Error típico	Estadístico *t*	Probabilidad	Inferior 95%	Superior 95%
Intercepción	–644 226.14	81 817.1131	–7.87397789	1.4123E-08	–811820.896	–476 631.383
Horas trabajadas	207.84	50.9105751	4.0824482	0.00033673	103.554201	312.12537
Salario diario	212.50	140.877135	1.50838955	0.14265674	–76.0761278	501.071324
Ingreso anual promedio del cónyuge	–2.38	0.86101115	–2.7682518	0.00988111	–4.147197	–0.61979429
Ingreso anual promedio de otros miembros de la familia	3.98	2.59462527	1.53333749	0.13641538	–1.33641265	9.29328507
Edad del jefe de familia	3 613.63	939.533293	3.84619261	0.00063373	1 689.07932	5 538.17269
Promedio del número de dependientes económicos	–7 435.67	4 151.7251	–1.79098401	0.08411331	–15 940.0965	1 068.74997
Promedio del grado máximo de escolaridad	11 310.67	7 516.6181	1.50475459	0.14358512	–4 086.4284	26 707.7596

Nótese aquí de nuevo la notación científica en algunas de las cifras; por ejemplo, el "1.4123E-08" de la probabilidad del coeficiente de la "intercepción", quinta columna y octavo renglón de abajo hacia arriba, equivale a 0.000000014123 en notación decimal; es decir, el "–8" de la notación científica indica que se debe correr el punto decimal de "1.4123" ocho posiciones a la izquierda.

El modelo ajustado es:

$$y = -644\,226.14 + 207.84x_1 + 212.50x_2 - 2.38x_3 + 3.98x_4 + 3\,613.63x_5 - 7\,435.67x_6 + 11\,310.67x_7$$

En donde:

y es el promedio de activos bancarios de la familia,
x_1 son las horas trabajadas,
x_2 es el salario diario,
x_3 es el ingreso anual promedio del cónyuge,
x_4 es el ingreso anual promedio de otros miembros de la familia,
x_5 es la edad promedio del jefe de familia,
x_6 es el promedio del número de dependientes económicos,
x_7 es el promedio del grado máximo de escolaridad.

El modelo da cuenta de 92.27% de la variación, ya que el coeficiente de determinación ajustado es, precisamente, 0.92272778.

Ahora, en la tabla 14.5 se muestra la tabla de correlaciones:

Tabla 14.5 Tabla de correlaciones para el ejemplo 14.7

	y	x_1	x_2	x_3	x_4	x_5	x_6	x_7
Y	1							
x_1	0.727	1						
x_2	0.792	0.570	1					
x_3	0.264	0.122	0.554	1				
x_4	0.284	–0.242	0.058	–0.041	1			
x_5	0.403	–0.077	0.042	–0.015	0.775	1		
x_6	–0.509	–0.342	–0.595	–0.692	0.050	–0.048	1	
x_7	0.648	0.691	0.886	0.538	–0.292	–0.326	–0.600	1

Puesto que la correlación entre las variables 3 y 4 con y es muy reducida (0.264 y 0.284), se les elimina del análisis. Además, como la variable 7 está altamente correlacionada con la variable 2 (0.886), se elimina la 7, que es la que entre ambas tiene menor correlación con y. Entonces se vuelve a calcular el modelo de regresión tras la eliminación de esas variables y se obtiene con Excel un nuevo modelo.

	Coeficientes
Intercepción	–732 003.661
x_1	268.216875
x_2	376.370676
x_5	3 507.7669
X_7	–2 320.09512

$$y = -732\,003.661 + 268.22x_1 + 376.37x_2 + 3\,507.77x_5 - 2\,320.10x_6$$

El cual explica 90.11% de la variación, ya que con Excel se obtiene:

Estadísticas de la regresión	
Coeficiente de correlación múltiple	0.955186
Coeficiente de determinación R^2	0.912381
R^2 ajustado	0.901075
Error típico	11479.35
Observaciones	36

Entonces se tiene ya un modelo aceptable, en términos de las variables incluidas y en términos del nivel de explicación de la variación total. Ahora, se evalúa este modelo reducido mediante el análisis de varianza y la prueba F, donde las hipótesis son:

$H_0 : \beta_1 = \beta_2 = \beta_5 = \beta_6 = 0$, no existe relación lineal entre y y el conjunto de variables independientes.
H_1 : Por lo menos una de las betas es diferente de 0: sí existe relación lineal entre y y cuando menos una de las variables independientes.

En la tabla de resultados de Excel se obtuvo:

Análisis de varianza					
	Grados de libertad	Suma de cuadrados	Promedio de los cuadrados	F	Valor crítico de F
Regresión	4	42 537 777 618	1.06E+10	80.70119	6.17E-16
Residuos	31	4 085 042 325	1.32E+08		
Total	35	46 622 819 944			

La probabilidad de obtener el valor observado de F, 80.70 es de 0.000000000000000617, es decir, prácticamente 0, por lo que se rechaza la hipótesis nula y se concluye que por lo menos una de las betas es diferente de 0, lo cual, a su vez indica que sí existe relación lineal entre y y cuando menos una de las variables independientes.

Cuando, como en este caso, se rechaza la hipótesis nula y se concluye que sí existe una relación lineal entre la variable dependiente y cuando menos una de las independientes, lo que sigue es realizar pruebas sobre los coeficientes de cada una de las variables independientes incluidas en el modelo con el propósito de verificar cuál o cuáles de ellas son significativas. En la sección siguiente se revisa este procedimiento.

14.4.3 Inferencias sobre coeficientes de regresión parciales individuales

Los coeficientes de regresión parcial obtenidos mediante la muestra se utilizan para evaluar la importancia de las variables predictoras individuales. Se realizan pruebas en donde las hipótesis son:

$$H_0: \beta_i = 0$$
$$H_1: \beta_i \neq 0$$

La prueba para evaluar estos coeficientes se realiza utilizando la t de Student como estadístico de prueba:

$$t = \frac{b_i - \beta_i}{s_{b_i}} \tag{14.6}$$

con $n - k - 1$ grados de libertad.

en donde:

s_{b_i} es el error estándar de la i-ésima variable. Como el cálculo de estos errores estándar es considerablemente laborioso, su análisis detallado escapa al alcance de este texto y, por ello, se aprovechan los resultados de Excel para su cálculo.

El estadístico de prueba:

$$t = \frac{b_i - \beta_i}{s_{b_i}}$$

En donde s_{b_i} es el error estándar de la i-ésima variable independiente.

Ejemplo 14.8

En el ejemplo 14.7 se analizó un modelo inicial con 7 variables y se utilizaron las correlaciones entre cada par de ellas para eliminar las que no eran útiles y se llegó a un modelo con sólo 4 variables:

$$y = -732\,003.661 + 268.22x_1 + 376.37x_2$$
$$+ 3\,507.77x_5 - 2\,320.10x_6$$

En donde:

x_1 son las horas trabajadas,
x_2 es el salario diario,
x_5 es la edad promedio del jefe de familia,
x_6 es el promedio del número de dependientes económicos.

Pruebe la significación de los coeficientes de cada una de estas variables, con un nivel de significación de 1 por ciento.

Solución: Las hipótesis son:

$$H_0: \beta_1 = 0$$
$$H_0: \beta_2 = 0$$
$$H_0: \beta_5 = 0$$
$$H_0: \beta_6 = 0$$

Contra las hipótesis alternativas:

$$H_1: \beta_1 \neq 0$$
$$H_1: \beta_2 \neq 0$$
$$H_1: \beta_5 \neq 0$$
$$H_1: \beta_6 \neq 0$$

Se reproducen en la tabla 14.6, los resultados de los últimos renglones de la tabla de resultados de Excel para el modelo

$$\hat{y} = -732\,003.661 + 268.22x_1 + 376.37x_2 + 3\,507.77x_5 - 2\,320.10x_6$$

los cuales incluyen los valores del estadístico t para cada una de las variables del modelo.

Tabla 14.6 Datos para el ejemplo 14.8

	Coeficientes	Error típico	Estadístico t	Probabilidad	Inferior 95%	Superior 95%
Intercepción	–732 003.661	78 761.9388	–9.29387559	1.788E-10	–892 639.694	–571 367.628
Horas trabajadas	268.216875	37.2055025	7.20906471	4.1646E-08	192.335753	344.097998
Salario diario	376.370676	59.2652753	6.35061044	4.5428E-07	255.498351	497.243002
Edad del jefe de familia	3 507.7669	451.532191	7.76858655	9.1474E-09	2 586.86093	4 428.67288
Promedio del número de dependientes económicos	–2 320.09512	3697.9515	–0.62740009	0.53499131	–9 862.1169	5 221.92666

En la cuarta columna de esta tabla se anotan los valores calculados del estadístico t de Student para cada variable, según la fórmula y en la quinta columna se anota su correspondiente probabilidad.

Como el nivel de significación se estableció en 0.01 se rechaza la hipótesis nula si el valor de la probabilidad de t, que es el que se anota en la quinta columna, es inferior. Así, se observa que las variables que implican un rechazo de la hipótesis nula son: horas trabajadas, salario diario y edad del jefe de familia, las cuales serían, entonces, las únicas variables significativas para el modelo.

En conclusión se puede decir que, con los datos muestrales con los que se cuenta, con un nivel de significación de 1% (es decir, con una probabilidad de rechazar equivocadamente la hipótesis nula), que los coeficientes de las variables que son diferentes de 0 son horas trabajadas, salario diario y edad del jefe de familia.

Al llegar a este punto, conviene volver a calcular el modelo de regresión lineal múltiple, utilizando solamente las variables cuyos coeficientes resultaron significativos.

Ejemplo 14.9

En la tabla 14.7 se muestran los resultados obtenidos aplicando la regresión lineal múltiple a los datos de "Promedio de activos bancarios de la familia" como variable dependiente y "horas trabajadas", "salario diario" y "edad del jefe de familia" como variables independientes.

Tabla 14.7 Resultados de regresión lineal múltiple para el ejemplo 14.9

Estadísticas de la regresión						
Coeficiente de correlación múltiple			0.95460384			
Coeficiente de determinación R^2			0.91126849			
R^2 ajustado			0.90294991			
Error típico			11 370.0721			
Observaciones			36			
Análisis de varianza						
	Grados de libertad	**Suma de cuadrados**	**Promedio de los cuadrados**	***F***	**Valor crítico de *F***	
Regresión	3	4.2486E+10	1.4162E+10	109.546171	6.5282E-17	
Residuos	32	4 136 913 254	129 278 539			
Total	35	4.6623E+10				
	Coeficientes	**Error típico**	**Estadístico *t***	**Probabilidad**	**Inferior 95%**	**Superior 95%**
Intercepción	–743 734.286	75 782.1403	–9.81411032	3.5808E-11	–898 097.454	–589 371.118
Horas trabajadas	268.385786	36.8503476	7.28312767	2.813E-08	193.324085	343.447488
Salario diario	395.493111	50.3434131	7.85590581	5.8038E-09	292.946935	498.039287
Edad del jefe de familia	3 516.17532	447.036637	7.86551934	5.6539E-09	2 605.59149	4 426.75915

Ahora el modelo es:

$$\hat{y} = -743\,734.286 + 268.39x_1 + 395.49x_2 + 3\,516.18x_5$$

Se observa en la tabla que ahora todos los coeficientes de las variables son estadísticamente significativos, pues su probabilidad de ocurrencia es ínfima.

Además, en este modelo final, el coeficiente de determinación ajustado, 0.90294991 indica que aproximadamente 90.29% de la variación en la variable dependiente queda explicada por el modelo. Comparando este coeficiente con los obtenidos para el modelo completo (7 variables independientes) y el de 4 variables independientes, 0.92272778 y 0.901075, se puede apreciar que la pérdida de capacidad explicativa respecto al modelo completo es muy reducida: 0.92272778 – 0.90294991 = 0.019778, es decir, de apenas 2%, en tanto que, por su parte, respecto al modelo con 4 variables independientes se dio, de hecho, un aumento en la capacidad de explicación:

$$0.90294991 - 0.901075 = 0.001875.$$

Se llegó, entonces, al mejor modelo que es posible construir con los datos de que se dispone.

Sin embargo, aún falta hacer algunas pruebas para verificar que el modelo es aceptable, y tienen que ver con los supuestos relacionados con los residuales, es decir, con las diferencias entre los valores muestrales de la variable dependiente y los valores estimados conseguidos a partir del modelo de regresión lineal múltiple. En la siguiente sección se analiza esto.

14.4.4 Análisis de residuales

Si el modelo es adecuado, los residuales que se obtienen con la ecuación de regresión múltiple ajustada con mínimos cuadrados se deben comportar de acuerdo con los supuestos en los que se basa el modelo, principalmente:

- Su suma debe ser igual a 0.
- Se deben distribuir en forma normal (de campana).

Los métodos para analizar residuales se pueden dividir en 2, los gráficos, que son los más sencillos y los que más se utilizan en la práctica y los analíticos. Se abordan aquí los métodos gráficos.

Ejemplo 14.10

Realice el análisis de residuales para el modelo "final", que es al que se llegó en el ejemplo 14.9 y que incluye sólo 3 variables independientes.

Solución: Excel permite hacer este análisis. Si se abre la ventana de "Regresión" de la sección de "Análisis de datos" de la pestaña de "Datos" de la cinta de opciones de Excel, aparece el cuadro de diálogo que se muestra en la figura 14.1.

En las secciones de la figura 14.1 se revisó lo que se debe anotar en las secciones "Entrada" y "Salida". Lo que resulta de interés ahora son las 2 secciones de abajo, "Residuales" y "Probabilidad normal". Se revisa en seguida el uso de estos 2 resultados y que se refiere, precisamente, al análisis de residuales que, como se dijo antes, sirve para evaluar la adecuación del modelo a sus supuestos.

En la tabla 14.8 se muestran los resultados numéricos obtenidos de Excel y que son los que se utilizaron en el análisis de regresión lineal múltiple hecho hasta aquí. Sin embargo, esta tabla tiene al final una sección (la más larga), dividida en 2 grupos de columnas, que tienen como encabezado "Análisis de los residuales" y "Resultados de datos de probabilidad".

Figura 14.1 Cuadro de diálogo para regresión.

Tabla 14.8 Tabla de Excel incluyendo análisis de residuales

Estadísticas de la regresión						
Coeficiente de correlación múltiple			0.95460384			
Coeficiente de determinación R^2			0.91126849			
R^2 ajustado			0.90294991			
Error típico			11 370.0721			
Observaciones			36			
Análisis de varianza						
	Grados de libertad	**Suma de cuadrados**	**Promedio de los cuadrados**	***F***	**Valor crítico de *F***	
Regresión	3	4.2486E+10	1.4162E+10	109.546171	6.5282E-17	
Residuos	32	4 136 913 254	129 278 539			
Total	35	4.6623E+10				

(*continúa*)

Tabla 14.8 (*continuación*)

	Coeficientes	Error típico	Estadístico *t*	Probabilidad	Inferior 95%	Superior 95%
Intercepción	–743 734.286	75 782.1403	–9.81411032	3.5808E-11	–898 097.454	–589 371.118
Horas trabajadas	268.385786	36.8503476	7.28312767	2.813E-08	193.324085	343.447488
Salario diario	395.493111	50.3434131	7.85590581	5.8038E-09	292.946935	498.039287
Edad del jefe de familia	3 516.17532	447.036637	7.86551934	5.6539E-09	2 605.59149	4 426.75915

Análisis de los residuales					Resultados de datos de probabilidad	
Observación	**Pronóstico promedio de activos bancarios de la familia**	**Residuos**	**Residuos estándares**		**Percentil**	**Promedio de activos bancarios de la familia**
1	–11 953.112	35 278.112	3.24489875		1.38888889	17 125
2	42 407.2313	–20 669.7313	–1.90121243		4.16666667	17 812.5
3	44 367.7446	–19 530.2446	–1.79640186		6.94444444	20400
4	49 132.3525	–14 057.3525	–1.29300245		9.72222222	21 737.5
5	43 616.7173	–5 266.71727	–0.48443534		12.5	23 325
6	63 550.5358	–8 063.03579	–0.74164215		15.2777778	24 837.5
7	20 391.19	–3 266.19003	–0.30042583		18.0555556	35 075
8	73 501.0659	–2 838.56591	–0.26109274		20.8333333	38 350
9	24 957.8815	–7 145.38152	–0.65723584		23.6111111	55 000
10	52 268.2415	2 731.75853	0.25126854		26.3888889	55 487.5
11	88 258.8872	6 203.61285	0.57061148		29.1666667	57 675
12	60 109.6225	–984.622514	–0.09056608		31.9444444	59 125
13	66 920.7231	–9 245.7231	–0.85042633		34.7222222	63 362.5
14	60 717.2265	2 645.27354	0.24331361		37.5	70 262.5
15	898.756781	19 501.2432	1.7937343		40.2777778	70 662.5
16	81 018.832	–7 181.33203	–0.66054259		43.0555556	73 837.5
17	95 021.0599	5 503.94007	0.50625522		45.8333333	84 862.5
18	142 266.426	16 608.5743	1.52766514		48.6111111	86 100
19	85 437.3541	5 187.64586	0.47716231		51.3888889	90 625
20	73 204.6374	–2 942.13739	–0.2706193		54.1666667	91 150
21	94 328.5122	983.987754	0.0905077		56.9444444	91 162.5
22	95 383.558	1 416.44201	0.13028506		59.7222222	91 562.5
23	122 049.059	5 938.44089	0.54622082		62.5	91 750
24	97 116.6538	120.846238	0.0111155		65.2777778	94 462.5
25	93 913.5459	–2 163.54585	–0.19900405		68.0555556	95 312.5
26	123 704.022	16 795.978	1.54490263		70.8333333	96 325
27	98 642.9307	–2 317.93069	–0.21320445		73.6111111	96 800
28	92 999.9383	–1 849.93826	–0.17015827		76.3888889	97 237.5
29	97 558.3737	–5 995.87366	–0.55150352		79.1666667	98 562.5
30	96 501.3529	–1 0401.3529	–0.95672174		81.9444444	100 525
31	118 373.541	1 2251.4592	1.12689547		84.7222222	103 962.5
32	102 405.464	–3 842.96401	–0.35347779		87.5	116 725
33	77 662.1527	7 200.34731	0.66229162		90.2777778	127 987.5
34	113 957.02	2 767.9799	0.2546002		93.0555556	130 625

Observación	Pronóstico promedio de activos bancarios de la familia	Residuos	Residuos estándares		Percentil	Promedio de activos bancarios de la familia
35	94 781.5456	–3 619.04558	–0.33288166		95.8333333	140 500
36	113 716.457	–9 753.95693	–0.89717394		98.6111111	158 875

Comenzando con el "Análisis de los residuales", este grupo de 4 columnas contiene, en la primera, la simple numeración de las observaciones, de los datos. La segunda columna "Pronóstico promedio de activos bancarios de la familia" son, precisamente los pronósticos realizados a través de la ecuación de regresión del modelo:

$$y = -743\,734.286 + 268.39x_1 + 395.49x_2 + 3\,516.18x_5$$

Por ejemplo, el primer pronóstico, –11 953.112 se obtuvo sustituyendo en esta ecuación los valores que tienen las variables en la primera observación de los datos originales. Se pueden ver en la tabla 14.3 los valores de (todas) las variables para cada una de las observaciones, o sea los grupos demográficos en los que se dividieron las familias del estudio. Se reproducen aquí, como la tabla 14.9, los encabezados de aquella tabla, junto con los datos para las 2 primeras observaciones.

Tabla 14.9 Las 2 primeras observaciones del conjunto de datos para el ejemplo de los 6 000 hogares divididos en 36 grupos demográficos

Grupo demográfico	Promedio de activos bancarios de la familia	Horas trabajadas	Salario diario	Ingreso anual promedio del cónyuge	Ingreso anual promedio de otros miembros de la familia	Edad del jefe de familia	Promedio del número de dependientes económicos	Promedio del grado máximo de escolaridad
1	23 325.00	1 985	142.30	6 912.50	4 762.50	40.60	3.833	6.6
2	21 737.50	2 042	230.40	13 562.50	4 100.00	41.80	2.444	8.2

Puede verse aquí que los valores en la primera observación (el primer grupo demográfico), para las variables que finalmente forman parte del modelo, x_1, x_2 y x_5 son, respectivamente, 1985, 142.30 y 40.60, para horas trabajadas, salario diario y la edad del jefe de familia. Sustituyendo estos valores en el modelo de regresión se tiene:

$$y = -\,743\,734.286 + 268.39(1985) + 395.49(142.3) + 3\,516.18(40.6)$$

Cuyo resultado es, precisamente, –11 953.112, con una pequeña diferencia debida al redondeo, ya que no se anotaron completos los coeficientes del modelo, tal como aparecen en la tabla 14.8.

Por supuesto, todos los demás residuales se obtuvieron de la misma manera y es claro aquí, de nuevo, la gran cantidad de labor que el paquete Excel ahorra con los cálculos.

En la tercera columna aparecen los "Residuos", que son la diferencia entre el valor de y en los datos originales y el pronóstico a partir de la regresión. Así, el primer residuo estándar es la diferencia entre el primer valor de y en los datos originales, 23 325, como puede verse en la tabla 14.9 y el primer pronóstico, –11 953.112: 23 325 – (– 11 953.112) = 23 325 + 11 953.112 = 35 278.112. Y así con los demás residuos.

La última columna de este grupo, titulada "Residuos estándares", a los que también se suele denominar *residuales estandarizados* se obtienen dividiendo cada residual entre la desviación estándar de los residuales. Aunque Excel no la muestra, esta desviación estándar se puede calcular fácilmente con la función de la desviación estándar muestral de Excel "DESVEST". Anotando en cualquier celda "=DesvEst(rango de los residuales)" se obtiene como resultado 10 871.86836 y dividendo, por ejemplo, el primer residual, 35 278.112 entre esa desviación estándar se obtiene 3.24489875, que es el primer residual estándar o estandarizado. Comienza ahora el análisis de estos residuales:

1. En primer lugar, se verifica que su suma sea 0, cosa que se puede hacer fácilmente con Excel para ver que la suma es, efectivamente, 0 (o casi, por cuestiones de redondeo): 0.000000000742. Esta verificación es simplemente una prueba de que no se cometieron errores en los cálculos; por supuesto con Excel esto no ocurre.
2. Una segunda prueba visual sobre los residuales consiste en dibujar un histograma de los residuales, e_i, el cual debe tener una forma aproximadamente normal para verificar el supuesto de normalidad.

   En la tabla 14.10 se ha construido una serie de clases y frecuencias con los datos de los residuales, con la función de Excel "Frecuencia", la cual se analizó con detalle en el capítulo 2, en la sección 2.2.5.

Tabla 14.10 Distribución de frecuencias de los residuales del ejemplo 14.10

Clases (residuales)	Frecuencias f
De –25 000 a menos de –20 000	1
De –20 000 a menos de –15 000	1
De –15 000 a menos de –10 000	2
De –10 000 a menos de –5 000	7
De – 5 000 a menos de 0	9
De 0 a menos de 5 000	6
De 5 000 a menos de 10 000	5
De 10 000 a menos de 15 000	1
De 15 000 a menos de 20 000	3
De 20 000 a menos de 25 000	0
De 25 000 a menos de 30 000	0
De 30 000 a menos de 35 000	0
De 35 000 a menos de 40 000	1
Suman las frecuencias	36

En la figura 14.2 se muestra el histograma de esa distribución de frecuencias de los residuales.

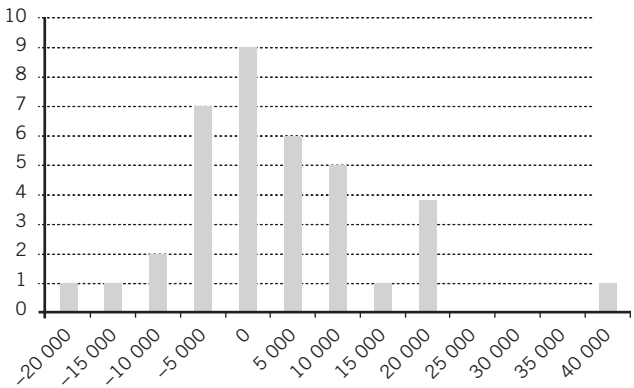


Figura 14.2 Histograma de la distribución de frecuencias de los residuales del ejemplo 14.10.

Se aprecia que la distribución de los residuales tiene forma aproximadamente normal, exceptuando por la observación del extremo derecho, con lo cual se puede considerar que el modelo cumple aproximadamente con el supuesto de normalidad de los residuales.

Existen técnicas estadísticas para abordar el tema de observaciones que se desvían notablemente del resto, como es el caso de esa observación del extremo derecho de la figura 14.2, a las que se denomina *observaciones aberrantes* basadas en su eliminación. Sin embargo, escapa a los alcances de este texto.

3. La gráfica de los residuales sirve para verificar que la relación entre cada variable independiente con la dependiente es lineal. Se busca observar en los residuales que las cantidades de residuales positivos y negativos son aproximadamente iguales y que se distribuyan en forma aleatoria, es decir, que no se observen patrones en el diagrama de dispersión.

   Para revisar que aproximadamente la mitad de los residuales sean positivos y la mitad negativos, se puede utilizar la función de Excel, la función "Si", que es una función condicional. Anotando en la celda contigua al primer residual la función "=SI(A1>0,1,)", el valor que Excel anota en la celda es "1" si el residual es mayor que 0 y anota 0 si esto no se cumple. Corriendo esta fórmula hacia abajo hasta el renglón 36 aparecería 1 en todas las celdas cuyo residual contiguo sea mayor que 0 y, para terminar, simplemente se suman todos ellos de la columna contigua a los residuales, con lo que se obtiene que 16 de ellos son positivos, es decir, poco menos de la mitad, con lo que se verifica que estos residuales cumplen también con este criterio.

   Esta distribución de los residuales, la mitad positivos y la mitad negativos, se puede apreciar visualmente si se les grafica como diagrama de dispersión. Se hace esto en la figura 14.3.

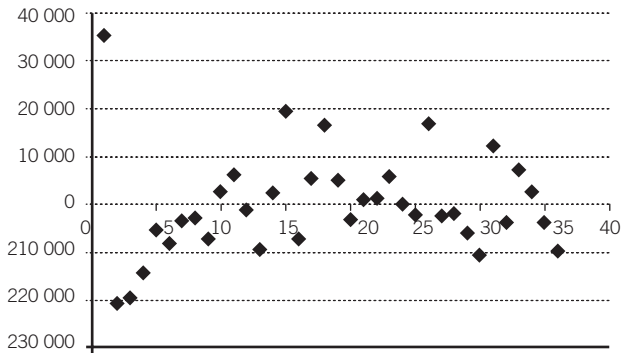


Figura 14.3 Diagrama de dispersión de los residuales del ejemplo 14.10.

En la figura 14.3 se puede ver que los puntos no parecen seguir un patrón, por lo que se concluye que sí se puede considerar que la relación entre las variables independientes con la dependiente es lineal.

4. Una cuarta prueba de cumplimiento con los supuestos del modelo, con base en el análisis de los residuales, lo proporciona Excel, a partir de un "Gráfico de probabilidad normal" que genera como parte del mecanismo de "Regresión" y que permite evaluar si la distribución de esos residuales es normal. Se revisa esto en seguida pero, antes, se debe mencionar que "Regresión" de Excel produce 7 gráficas:
   - Tres en las que se grafica cada variable independiente contra los residuales.
   - Tres en las que se grafica cada variable contra los valores estimados con la recta de regresión.
   - Una, llamada *gráfico de probabilidad normal*, mencionada antes y que se muestra en seguida como figura 14.4.

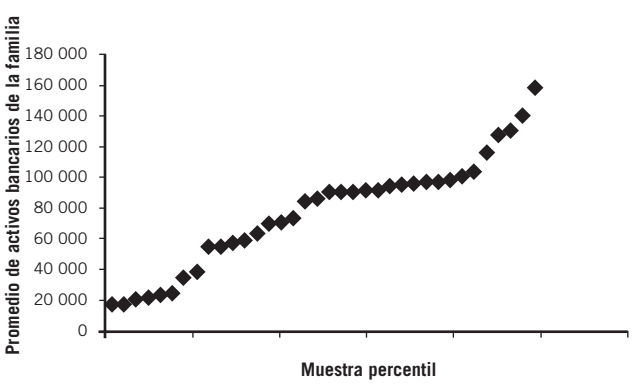


Figura 14.4 Gráfica de probabilidad normal para los residuales estandarizados del ejemplo 14.10.

Sobre estas gráficas, es importante comentar que las 6 primeras no suelen ser parte de análisis básicos de residuales, por lo que no se revisan aquí.

Respecto a la gráfica de probabilidad normal, se trata de la gráfica de los "Resultados de datos de probabilidad" de los resultados de Excel, anotados en la parte inferior, en las 2 últimas columnas, que es el segundo grupo de datos que Excel arroja en la sección de "Análisis de residuales" y se utiliza, como ya se mencionó, para evaluar si la distribución de esos residuales es normal.

En la tabla 14.8 se observa que los encabezados de estas 2 columnas son "Percentil" y "Promedio de activos bancarios de la familia", es decir, la última columna contiene los datos originales de la variable y, sólo que en orden ascendente de magnitud, mientras que la columna de percentil se calcula de manera que sus valores partan (casi) de 0, 1.388888888 en este caso, y terminen en 100 (o casi).

Lo que hace Excel es que primero divide 100 (el uso del 100 es para asegurar que los percentiles estén entre 0 y 100) entre el número de datos, lo cual da como resultado 2.77777778; la mitad de este valor es, precisamente, el valor inicial de estos percentiles, 1.388888888. A partir del segundo dato, lo que se hace es ir sumando 2.77777778 a cada valor dado para obtener el siguiente. Por ejemplo, el segundo valor es la suma del primero, 1.388888888, más 2.77777778, lo cual suma 4.166666667, si a este último valor se le suma otra vez 2.77777778, se obtiene el tercero, 9.72222222, y así sucesivamente. Lo que este conjunto de percentiles representa gráficamente es una recta con pendiente de 45°, como se ve en la figura 14.5.

Así, el propósito de graficar los valores muestrales de la variable, ordenados de menor a mayor y utilizando estos percentiles como los valores del eje horizontal es ver qué tanto se ajustan esos valores a un recta con pendiente de 45°, lo cual es de esperar en una serie con distribución aproximadamente normal, cosa que sí sucede en la gráfica que se presenta en la figura 14.5, que es la gráfica que Excel genera con esos valores percentiles y los valores muestrales de la variable dependiente, lo cual confirma que se puede considerar que los residuales tienen un distribución aproximadamente normal.

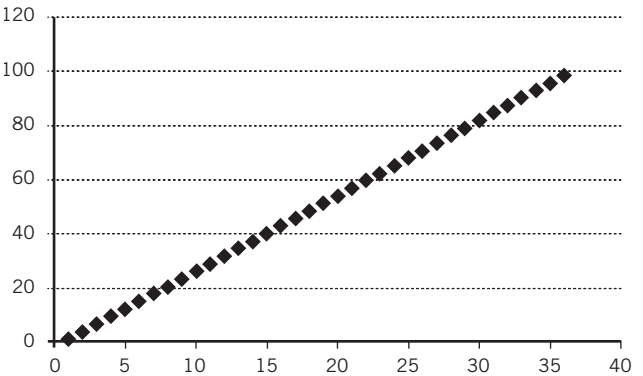


Figura 14.5 Gráfica de los percentiles del ejemplo 14.10.

Una nota importante antes de terminar: se incluye aquí la gráfica de los percentiles sólo para mostrar que es una recta que asciende con inclinación de 45 grados, pero no forma parte del análisis de residuales. La gráfica que se debe analizar es la de la probabilidad normal.

ejercicios 14.4 Evaluación de la ecuación de regresión

1. Retome los datos del ejercicio 14.3, punto 3, sobre PIB y distribución de las poblaciones en las actividades económicas primaria, secundaria y terciaria para 45 países incluyendo México:

a) Obtenga el coeficiente de determinación múltiple e interprételo.

b) Plantee la hipótesis para realizar una prueba global de significación para la regresión utilizando el estadístico F de Fisher, interprételo y establezca la conclusión, tanto en términos de aceptar o rechazar la hipótesis nula como en términos de los datos mismos.

c) Si las conclusiones de *b*) muestran que tiene sentido, plantee hipótesis para realizar pruebas sobre los coeficientes de regresión parcial de la ecuación de regresión. Determine los estadísticos de prueba t de Student y establezca las conclusiones correspondientes.

d) ¿Las conclusiones de *c*) indican que se debe modificar el modelo? Si la respuesta es afirmativa, hágalo y repita los incisos *c*) y *d*).

e) Realice un análisis de los residuales del modelo al que llegó hasta aquí:

i) ¿El histograma de los residuales tiene forma aproximadamente normal? Concluya con base en su respuesta.

ii) ¿Las cantidades de residuales positivos y negativos son aproximadamente iguales? Concluya con base en su respuesta.

iii) Analice el gráfico de probabilidad normal para esos residuales y fundamente la conclusión que se pueda extraer de esta gráfica.

2. Retome los datos del ejercicio 14.3, punto 4, sobre la empresa de paquetería que evalúa los tiempos de recorrido de sus rutas y responda lo siguiente:

a) Determine el coeficiente de determinación múltiple e interprételo.

b) Plantee la hipótesis para realizar una prueba global de significación para la regresión utilizando el estadístico F de Fisher, interprételo y establezca la conclusión,

tanto en términos de aceptar o rechazar la hipótesis nula como en términos de los datos mismos.

c) Si las conclusiones de *b*) muestran que tiene sentido, plantee hipótesis para realizar pruebas sobre los coeficientes de regresión parcial de la ecuación de regresión. Determine los estadísticos de prueba *t* de Student y establezca las conclusiones correspondientes.

d) ¿Las conclusiones de *c*) indican que se debe modificar el modelo? Si la respuesta es afirmativa, hágalo y repita los incisos *c*) y *d*).

e) Realice un análisis de los residuales del modelo al que se llegó hasta aquí:

i) ¿El histograma de los residuales tiene forma aproximadamente normal? Concluya con base en su respuesta.

ii) ¿Las cantidades de residuales positivos y negativos son aproximadamente iguales? Concluya con base en su respuesta.

iii) Analice el "Gráfico de probabilidad normal" para esos residuales y fundamente la conclusión que se pueda extraer de esta gráfica.

3. Retome los datos del ejercicio 14.3, punto 5, sobre características y precios de autos híbridos que se venden en Estados Unidos:

a) Determine el coeficiente de determinación múltiple e interprételo.

b) Plantee la hipótesis para realizar una prueba global de significación para la regresión utilizando el estadístico *F* de Fisher, interprételo y establezca la conclusión, tanto en términos de aceptar o rechazar la hipótesis nula como en términos de los datos mismos.

c) Si las conclusiones de *b*) muestran que tiene sentido, plantee hipótesis para realizar pruebas sobre los coeficientes de regresión parcial de la ecuación de regresión. Determine los estadísticos de prueba *t* de Student y establezca las conclusiones correspondientes.

d) ¿Las conclusiones de *c*) indican que se debe modificar el modelo? Si la respuesta es afirmativa, hágalo y repita los incisos *c*) y *d*).

e) Realice un análisis de los residuales del modelo al que se llegó hasta aquí:

i) ¿El histograma de los residuales tiene forma aproximadamente normal? Concluya con base en su respuesta.

ii) ¿Las cantidades de residuales positivos y negativos son aproximadamente iguales? Concluya con base en su respuesta.

iii) Analice el gráfico de probabilidad normal para esos residuales y fundamente la conclusión que se pueda extraer de esta gráfica.

4. En la tabla siguiente se muestran las calificaciones que obtuvieron 15 estudiantes en 4 asignaturas.

	Estadística inferencial	Estadística descriptiva	Fundamentos de matemáticas	Matemáticas financieras
1	6.5	7.5	7.5	6.5
2	6.5	7	6.5	6
3	6.5	8	6.5	7
4	7.5	10	7	8
5	8	9	9	8.5
6	7.5	8	8.5	7.5
7	6	6.5	7.5	5.5
8	6.5	8	7	7.5
9	6.5	7.5	7	6.5
10	6.5	7	6.5	5
11	5.5	6	6	5
12	7	8	5.5	7
13	7.5	8.5	8	7.5
14	7.5	9.5	7	8
15	7.5	6.5	7.5	5

a) Construya la ecuación regresión lineal múltiple, utilizando "estadística inferencial" como la variable dependiente.

b) Determine el coeficiente de determinación múltiple e interprételo.

c) Plantee la hipótesis para realizar una prueba global de significación para la regresión utilizando el estadístico *F* de Fisher, interprételo y establezca la conclusión, tanto en términos de aceptar o rechazar la hipótesis nula como en términos de los datos mismos.

d) Si las conclusiones de *b*) muestran que tiene sentido, plantee hipótesis para realizar pruebas sobre los coeficientes de regresión parcial de la ecuación de regresión. Determine los estadísticos de prueba *t* de Student y establezca las conclusiones correspondientes.

e) ¿Las conclusiones de *c*) indican que se debe modificar el modelo? Si es afirmativo, hágalo y repita los incisos *c*) y *d*).

f) Realice un análisis de los residuales del modelo al que se llegó hasta aquí:

i) ¿El histograma de los residuales tiene forma aproximadamente normal? Concluya con base en su respuesta.

ii) ¿Las cantidades de residuales positivos y negativos son aproximadamente iguales? Concluya con base en su respuesta.

iii) Analice el gráfico de probabilidad normal para esos residuales y fundamente la conclusión que se pueda extraer de esta gráfica.

5. En un estudio de gastos con tarjeta de crédito se reunieron datos sobre ingreso anual, número de miembros de la familia y gasto anual con tarjeta de crédito y se obtuvieron los siguientes resultados:

Ingreso anual	Miembros de la familia	Gastos anuales con tarjeta
540 000	4	4 020
300 000	3	3 160
320 000	5	5 100
500 000	6	4 750
310 000	3	1 870
550 000	3	4 100
370 000	2	2 750
400 000	3	3 350
660 000	5	4 770
510 000	4	4 150
250 000	4	4 210
480 000	5	4 220
270 000	2	2 480
330 000	3	2 520
650 000	4	4 220
630 000	5	4 970
420 000	7	4 420
210 000	3	2 450
440 000	2	3 000
370 000	6	4 180
620 000	7	5 680
210 000	4	3 630
550 000	8	5 310
420 000	3	3 020
410 000	8	4 830
340 000	6	3 590
670 000	5	5 040
500 000	3	3 610
670 000	6	5 350
550 000	7	5 370
520 000	3	3 890
620 000	4	4 710
640 000	3	4 160
220 000	4	3 580
290 000	5	3 890
390 000	3	3 980
360 000	2	3 130

Ingreso anual	Miembros de la familia	Gastos anuales con tarjeta
390 000	5	4 190
540 000	4	3 830
230 000	7	4 130
270 000	3	2 920
260 000	8	4 610
610 000	3	4 270
300 000	8	3 070
220 000	3	3 080
460 000	3	4 820
660 000	5	5 150

a) Construya la ecuación regresión lineal múltiple, utilizando "gastos anuales con tarjeta" como la variable dependiente.

b) Determine el coeficiente de determinación múltiple e interprételo.

c) Plantee la hipótesis para realizar una prueba global de significación para la regresión utilizando el estadístico F de Fisher, interprételo y establezca la conclusión, tanto en términos de aceptar o rechazar la hipótesis nula como en términos de los datos mismos.

d) Si las conclusiones de *b*) muestran que tiene sentido, plantee hipótesis para realizar pruebas sobre los coeficientes de regresión parcial de la ecuación de regresión. Determine los estadísticos de prueba t de Student y establezca las conclusiones correspondientes.

e) ¿Las conclusiones de *c*) indican que se debe modificar el modelo? Si es afirmativo, hágalo y repita los incisos *c*) y *d*).

f) Realice un análisis de los residuales del modelo al que se llegó hasta aquí:

i) ¿El histograma de los residuales tiene forma aproximadamente normal? Concluya con base en su respuesta.

ii) ¿Las cantidades de residuales positivos y negativos son aproximadamente iguales? Concluya con base en su respuesta.

iii) Analice el "Gráfico de probabilidad normal" para esos residuales y fundamente la conclusión que se pueda extraer de esta gráfica.

14.5 Uso del modelo de regresión lineal múltiple

Una vez que se verificó que se tiene un buen modelo de regresión lineal simple, se puede utilizar la ecuación de regresión en 2 aplicaciones:

1. Para pronosticar el valor que es probable que asuma la variable dependiente para determinados valores de las variables independientes.

2. Para estimar el promedio de la subpoblación de valores y que se asume existen para determinada combinación de las variables independientes.

14.5.1 Intervalos de confianza para los pronósticos

Se asume de la siguiente manera el procedimiento para realizar pronósticos para valores probables de la variable dependiente:

$$\hat{y} \pm t_{\left(1-\frac{\alpha}{2}, n-k-1\right)} s_{\hat{y}_i} \tag{14.7}$$

En donde:

$\hat{y}$ es un valor puntual estimado de la variable dependiente, a partir de un conjunto de valores especificados de las variables independientes, $t_{\left(1-\frac{\alpha}{2}, n-k-1\right)}$ es el valor crítico de la distribución t de Student,

$s_{\hat{y}_i}$ es el error estándar de los valores estimados de la variable dependiente.

El procedimiento para calcular este error estándar es considerablemente laborioso y no se aborda aquí. Sin embargo, Excel lo calcula.

En el ejemplo 14.9 se aplicó la regresión lineal múltiple a los datos de promedio de activos bancarios de la familia como variable dependiente y horas trabajadas, salario diario y edad del jefe de familia como variables independientes y se obtuvo el modelo:

$$\hat{y} = -743\,743.286 + 268.39x_1 + 395.49x_2 + 3\,516.18x_5$$

A su vez, en la tabla 14.7 se presentaron los datos de los resultados de Excel para regresión lineal múltiple aplicada a esos datos. En esa tabla, en el quinto renglón se consigna el "error típico" que, en la terminología que se usa aquí (y que es la más común) es, precisamente, el error estándar de $\hat{y}$, y es igual a 11 370.0721.

Con esta información se pueden construir intervalos de confianza para pronósticos de $\hat{y}$, utilizando diferentes conjuntos de valores para las variables independientes.

EJEMPLO 14.11

Construya un intervalo de confianza de 90% para pronosticar el valor de $\hat{y}$, utilizando los siguientes valores para las variables independientes:

x_1	Horas trabajadas	2 200
x_2	Salario diario	300
x_5	Edad del jefe de familia	45

Solución: Sustituyendo valores en la ecuación de regresión:

$$\hat{y} = -743\,743.286 + 268.39x_1 + 395.49x_2 + 3\,516.18x_5$$

$$\hat{y} = -743\,743.286 + 268.39(2\,200) + 395.49(300) + 3\,516.18(45)$$

$$\hat{y} = -743\,743.286 + 590\,458 + 118\,647 + 158\,228.1$$

$$\hat{y} = 123\,598.81$$

Que es entonces la estimación puntual de la variable dependiente para ese conjunto de valores de las variables independientes.

Ahora el intervalo:

$$\hat{y} \pm t_{\left(1-\frac{\alpha}{2}, n-k-1\right)} s_{\hat{y}_i}.$$

Aquí, $n = 36$, $k = 3$, por lo que los grados de libertad son:

$$n - k - 1 = 36 - 3 - 1 = 32$$

Entonces el valor del estadístico t de Student es 2.037, ya que

$$P(-2.037 \le t \le 2.037 \mid gl = 32) = 0.90$$

En donde el valor de la t se obtuvo con la herramienta Excel, mediante "=DISTR.T.INV(0.05,32)"

$$123\,598.81 \pm 2.037(11\,370.0721) = 123\,598.81 \pm 23\,160.84$$

Es decir que se pronostica, con 90% de confianza de estar en lo correcto, que el valor que es probable que asuma el promedio de activos bancarios de la familia para ese conjunto de valores de las variables independientes está entre \$100 437.97 y \$146 759.65.

14.5.2 Intervalos de confianza para estimaciones de la media de una subpoblación de valores *y*

Estos intervalos se construyen de la siguiente manera:

$$\hat{\mu}_y \pm t_{\left(1-\frac{\alpha}{2}, n-k-1\right)} s_{\mu} \tag{14.8}$$

En donde:

$\hat{y}$ es la estimación puntual,

$t_{\left(1-\frac{\alpha}{2}, n-k-1\right)}$ es el estadístico de prueba t de Student,

s_{μ} es el error estándar de la estimación.

Cuando se cumplen las suposiciones básicas del modelo se pueden construir intervalos de predicción e intervalos de confianza. Sin embargo, al igual que sucede con el error estándar para los intervalos de pronóstico, el procedimiento para calcular el error estándar para realizar estimaciones es considerablemente laborioso y no se aborda aquí y, como no lo calcula Excel, se detiene aquí el análisis.

ejercicios 14.5 Uso del modelo de regresión lineal múltiple

1. Construya un intervalo de confianza de 95% para pronosticar el valor de $\hat{y}$, el PIB, del ejercicio 14.3, punto 3, utilizando los siguientes valores para las variables independientes, en caso de que hayan quedado incluidas en el modelo final:

% de población urbana	70
% en actividades económicas primarias	60
% en actividades económicas secundarias	30
% en actividades económicas terciarias	10

2. Haga un pronóstico por intervalo, con nivel de confianza de 98%, para el valor de $\hat{y}$, tiempo de recorrido para el modelo del ejercicio 4, de los ejercicios por sección 14.3, utilizando los siguientes valores para las variables independientes:

Recorrido en km	150
Núm. de paquetes entregados	3

3. Haga un pronóstico por intervalo, con nivel de confianza de 95%, para el valor de $\hat{y}$, precio del automóvil, del ejercicio 5, de los ejercicios por sección 14.3, utilizando los siguientes valores para las variables independientes:

Millas por galón en ciudad	55
Millas por galón en carretera	50
Potencia (hp)	150
Revoluciones por minuto	5 500

4. Construya un intervalo de confianza de 90% para pronosticar el valor de $\hat{y}$, la calificación en estadística inferencial, del ejercicio 14.4, punto 4, utilizando los siguientes valores para las variables independientes, en caso de que quedaran incluidas en el modelo final:

Estadística descriptiva	9
Fundamentos de matemáticas	9
Matemáticas financieras	9

5. Haga un pronóstico por intervalo, con nivel de confianza de 95%, para el valor de $\hat{y}$, gasto anual con tarjeta de crédito, del ejercicio 14.4, punto 5, utilizando los siguientes valores para las variables independientes:

Ingreso anual	250 000
Núm. de miembros de la familia	5

14.6 Variables independientes cualitativas

Los ejemplos revisados hasta aquí involucran variables numéricas. La metodología del análisis de regresión lineal múltiple permite utilizar también variables nominales, es decir, categóricas. Cuando se desea introducir variables categóricas en el análisis se utilizan variables ficticias dicotómicas que son variables que asumen sólo 2 valores, 0 y 1, y que se utilizan para identificar las diferentes categorías de una variable cualitativa. Algunos ejemplos de variables cualitativas y variables ficticias que se utilizan para cuantificarlas:

Variable cualitativa	Núm. de categorías	Variable ficticia
Sexo	2	$x = 1$ para femenino $x = 0$ para masculino
Ubicación	3	$x_1 = 1$ para urbana $x_1 = 0$ para rural o suburbana $x_2 = 1$ para rural $x_2 = 0$ para urbana o suburbana

(*continúa*)

(*continuación*)

Variable cualitativa	Núm. de categorías	Variable ficticia
Estado civil: soltero, casado, viudo, divorciado	4	$x_1 = 1$ para soltero $x_1 = 0$ para los otros casos $x_2 = 1$ para casado $x_2 = 0$ para los otros casos $x_3 = 1$ para casado $x_3 = 0$ para los otros casos

Nótese cómo, para representar a una variable categórica, se requiere una cantidad de variables ficticias igual al número de categorías menos 1. Por ejemplo, en el caso de la ubicación, se identifica la ubicación suburbana cuando tanto x_1 como x_2 son iguales a 0.

Ejemplo 14.12

Una agencia automotriz grande desea estimar los efectos que se observan sobre el tiempo de servicio a los automóviles que llegan al taller a partir de 3 variables independientes: los kilómetros recorridos desde el último servicio, el tipo de servicio (normal o express) y el mecánico que realiza la tarea. En la tabla 14.11 se muestran los datos de los que se dispone. Construya el modelo de regresión lineal múltiple que describa esa relación.

Tabla 14.11 Datos para el ejemplo 14.12

Tiempo de servicio (horas)	Km recorridos desde el último servicio (miles)	Tipo de servicio	Mecánico
3.0	5	normal	Ramón
3.1	9	express	Ramón
4.9	11	normal	Javier
1.9	6	express	Javier
3.0	5	normal	Javier
5.0	10	normal	Luis
4.3	12	express	Luis
4.9	11	express	Luis
4.5	7	normal	Ramón
4.6	9	normal	Javier
4.7	10	express	Luis

Solución: En primer lugar, como hay una sola variable independiente numérica, a ésta se le identifica como x_1. Como se tienen 2 variables categóricas, una dicotómica y otra con 3 posibles valores, se requieren las siguientes variables ficticias:

Variable cualitativa	Núm. de categorías	Variable ficticia
Tipo de servicio	2	$x_2 = 1$ para normal $x_2 = 0$ para expres
Mecánico	3	$x_3 = 1$ Ramón $x_3 = 0$ para Javier o Luis $x_4 = 1$ para Javier $x_4 = 0$ para Luis o Ramón

De esta manera se pueden modelar estos datos con la variable dependiente y con 4 variables independientes y el conjunto de datos es ahora:

Tiempo de servicio (horas) y	Km recorridos desde el último servicio (miles) x_1	Tipo de servicio x_2	Mecánico x_3	Mecánico x_4
3.0	5	1	1	0
3.1	9	0	1	0
4.9	11	1	0	1
1.9	6	0	0	1
3.0	5	1	0	1
5.0	10	1	0	0
4.3	12	0	0	0
4.9	11	0	0	0
4.5	7	1	1	0
4.6	9	1	0	1
4.7	10	0	0	0

Aplicando ahora la regresión lineal múltiple de Excel a estos datos, se obtienen los resultados que se muestran en la tabla 14.12.

Tabla 14.12 Resultados para el ejemplo 14.12

Estadísticas de la regresión				
Coeficiente de correlación múltiple	0.93106603			
Coeficiente de determinación R^2	0.86688395			
R^2 ajustado	0.77813992			
Error típico	0.49466302			
Observaciones	11			

(*continúa*)

(*continuación*)

Análisis de varianza						
	Grados de libertad	Suma de cuadrados	Promedio de los cuadrados	*F*	Valor crítico de *F*	
Regresión	4	9.5609419	2.39023547	9.76836326	0.00849322	
Residuos	6	1.46814901	0.2446915			
Total	10	11.0290909				
	Coeficientes	Error típico	Estadístico *t*	Probabilidad	Inferior 95%	Superior 95%
Intercepción	0.91618161	0.97087037	0.94367038	0.38176821	–1.45945259	3.2918158
Km recorridos desde el último servicio (miles)	0.3249709	0.08605964	3.77611257	0.00922232	0.11439054	0.53555125
Tipo de servicio	1.26152503	0.33839652	3.72794918	0.00976027	0.43349858	2.08955148
Mecánico	–0.49866123	0.50604916	–0.98540078	0.36248072	–1.73691892	0.73959645
Mecánico	–0.78084983	0.45538482	–1.71470323	0.1372249	–1.89513634	0.33343669

Así, el modelo de regresión lineal múltiple es:

$$\hat{y} = 0.916 + 0.325x_1 + 1.262x_2 - 0.499x_3 - 0.718x_4$$

Como la variable x_1 es numérica, su coeficiente no requiere mayor análisis. La variable x_2 puede tomar valores de 0 y 1, lo cual conduce a 2 ecuaciones de regresión diferentes: una para el servicio normal ($x_2 = 1$), otra para el servicio express ($x_2 = 0$):

$$x_2 = 1 \ : \ \hat{y} = 0.916 + 0.325x_1 + 1.262x_2 - 0.499x_3 - 0.718x_4$$
$$x_2 = 0 : \hat{y} = 0.916 + 0.325x_1 - 0.499x_3 - 0.781x_4$$

Observando las ecuaciones se puede ver que, como el coeficiente de esta variable x_2 es positivo, 1.262, el uso del servicio express elimina ese factor; de hecho, si las demás variables permanecen constantes, reduce el tiempo de servicio a los automóviles, lo cual confirma que ese servicio express es, como se anuncia, más veloz.

Por su parte, como los coeficientes de las otras 2 variables ficticias, x_3 y x_4 son negativas, significa que cuando x_3 es igual a 1, el mecánico del servicio es Ramón y señala que su actividad reduce el tiempo de servicio, al igual que sucede con Javier, ya que cuando x_4 es igual a 1, se mantiene la variable y eso hace que disminuya el tiempo de servicio.

Por su parte, cuando $x_3 = 0$, se trata de Javier o Luis, y cuando también es $x_4 = 0$, se trata de Luis o Ramón, lo cual quiere decir que el trabajo de Luis no contribuye a disminuir los tiempos de servicio de los autos.

ejercicios 14.6 Variables independientes cualitativas

1. En un estudio sobre el absentismo se recopilaron datos de 12 trabajadores, con los siguientes resultados:

Trabajador	Núm. de faltas	Edad	Sexo
1	5	26	H
2	0	31	H
3	1	63	M
4	6	34	H
5	7	46	M
6	11	28	M
7	5	56	H
8	10	42	H
9	3	23	H
10	2	59	M
11	1	32	H
12	6	27	H

Construya una ecuación de regresión lineal múltiple, con el número de faltas como variable dependiente y esbozar las conclusiones que se pueden extraer de ese modelo.

2. En un estudio sobre riesgos de ataques al corazón realizado durante varios años, se obtuvieron datos sobre edad, presión sanguínea y hábito de fumar. En la tabla que sigue se presentan los resultados para 20 sujetos.

% de riesgo	Edad	Presión	Fuma (1) No fuma (0)
13	56	151	0
25	66	162	0
14	57	154	0
57	85	176	1
29	58	195	0
52	75	188	1

(*continúa*)

(*continuación*)

% de riesgo	Edad	Presión	Fuma (1) No fuma (0)
19	55	154	1
32	77	119	0
37	79	134	1
16	77	97	0
23	70	151	0
37	69	172	1
16	66	134	1
49	76	208	1
16	59	198	0
37	81	118	1
9	65	165	0
35	79	124	1
4	61	116	0
38	58	206	1

Construya una ecuación de regresión lineal múltiple, con el porcentaje de riesgo como variable dependiente y esbozar las conclusiones que se pueden extraer de ese modelo.

3. Para evaluar la efectividad del programa de capacitación de una empresa, se tomó una muestra de 28 trabajadores con y sin capacitación y se determinó un índice basado en su experiencia y habilidad, así como el número de repeticiones de una tarea simple que lograban hacer antes de cometer un error. Los resultados son los siguientes:

Núm. de repeticiones	Capacitación / sin capacitación	Índice de experiencia y habilidad
86	C	82
56	C	207
80	C	107
29	S	194
55	C	132
30	S	144
35	S	175
95	C	20
30	S	137
92	C	72
69	S	132
32	S	166
68	C	220
82	C	83
53	S	157
55	C	206
50	C	131
80	S	95
80	C	95
67	S	83
86	S	82
71	S	96
85	C	85
31	S	240
32	C	145
55	C	246
55	S	97
92	C	85

Construya una ecuación de regresión lineal múltiple, con el número de repeticiones como variable dependiente y esboce las conclusiones que se pueden extraer de ese modelo.

4. Una agencia de bienes raíces que desea determinar si existe relación entre los precios de departamentos en venta y su superficie y la disponibilidad de cuarto de servicio, obtuvo los siguientes datos de una muestra de departamentos:

Precio	Superficie	Cuarto de servicio
126.6	120	Sí
116.1	102.6	No
113.55	87	No
128.85	105.6	Sí
118.65	115.8	No
105.6	72	Sí
113.7	93	Sí
128.85	115.8	Sí
117.75	95.4	Sí
118.8	90	Sí
130.05	114	Sí
118.95	83.4	Sí
111.75	92.4	No
125.7	113.4	Sí
115.2	95.4	No

Construya una ecuación de regresión lineal múltiple, con el precio como variable dependiente y plantee las conclusiones que se pueden extraer de ese modelo.

5. Un fabricante desea revisar la relación que tienen las ventas de su producto con el espacio de anaquel y la ubicación en los pasillos que los supermercados le otorgan. Para ello, tomó una muestra de 12 supermercados y obtuvo la información que se muestra en seguida:

Supermercado	Espacio en anaquel (m)	Ubicación en el pasillo	Ventas por semana (miles)
1	2	En medio	20.8
2	2	Al principio	28.6
3	2	En medio	18.2
4	3	En medio	24.7
5	3	En medio	31.2
6	3	Al principio	33.8
7	5	En medio	29.9
8	5	En medio	35.1
9	5	Al principio	36.4
10	7	En medio	33.8
11	7	En medio	37.7
12	7	Al principio	40.3

Construya una ecuación de regresión lineal múltiple, con las ventas semanales como variable dependiente y plantee las conclusiones que se pueden extraer de ese modelo.

14.7 Regresión por pasos

La técnica de regresión por pasos consiste en ir introduciendo o retirando variables independientes en el modelo de regresión, de una en una, hasta analizarlas todas y llegar a un modelo satisfactorio, de manera que son 2 los procedimientos que se pueden seguir:

1. Eliminación posterior.
2. Selección previa.

En las secciones siguientes se revisan ambos casos.

14.7.1 Eliminación posterior

En el ejemplo 14.1, que se refería a datos de las 9 principales compañías de 2011 que aparecen en el listado de las 500 empresas más importantes de México, que cada año publica la revista *Expansión*, se tenían inicialmente datos de ventas como variable dependiente y las variables independientes eran el activo, el pasivo, el patrimonio y el número de empleados. Se siguió un procedimiento de regresión por pasos de eliminación posterior que consistió en:

1. En primer lugar se determinó con Excel el modelo de regresión con todas las variables.
2. En seguida se calculó la matriz de correlación para encontrar los coeficientes de correlación entre cada par de variables con el objeto de:

   a) Determinar cuáles eran las variables independientes que tenían la correlación más alta con y, la variable dependiente, ya que lo apropiado sería incluir las variables independientes que tengan correlación alta con la variable dependiente.

   b) Identificar qué pares de variables estaban altamente correlacionadas entre sí, ya que no tiene sentido incluir en un modelo variables independientes que estén altamente correlacionadas entre sí, porque esa correlación es indicativa de que aportan la misma información al modelo; en el ejemplo se escogió, de entre las altamente correlacionadas entre sí, la que tenía la correlación más alta con y, y se eliminó la otra.

Se resumió lo anterior de la siguiente manera:

1. Ningún par de variables independientes debe tener una correlación inferior a -0.70 ni superior a 0.70.
2. Una variable independiente, a las que también se denominan *predictoras*, cuando se les utiliza para hace pronósticos sobre la variable dependiente, debe tener, preferentemente, una correlación fuerte con la variable dependiente.

Después de realizado lo anterior, se procedió a evaluar la ecuación de regresión,

a) En forma global, mediante:

- El coeficiente de determinación múltiple.
- El análisis de varianza y la prueba con el estadístico F de Fisher.

b) En forma específica se evaluaron los coeficientes de regresión individuales.

Aunque ya no fue con el modelo de las 500 empresas más importantes de México, más adelante se revisó un tercer tipo de evaluación de modelos de regresión lineal múltiple: la que se basa en el análisis de residuales, la cual se relaciona, más que con la adecuación en sí del modelo, con el cumplimiento de los supuestos en los que se basa.

En el análisis de los datos de las 500 empresas más importantes de México, la evaluación del modelo depurado con 2 variables independientes utilizando el análisis de varianza y la prueba con el estadístico F de Fisher, condujo a la conclusión de que se debía abandonar el análisis ya que, con los datos disponibles, no se podía obtener un buen modelo de regresión lineal múltiple.

Posteriormente, en el ejemplo 14.7 con datos de una muestra de 6 000 hogares con ingresos inferiores a $180 000 anuales, clasificando los hogares en 36 grupos demográficos con el propósito de evaluar los salarios diarios, se propuso hacer un análisis de regresión utilizando el promedio de activos bancarios de la familia como variable dependiente, y otras 7 variables como independientes. El procedimiento que se siguió en este caso fue:

1. Se realizó el análisis de regresión incluyendo todas las variables y se encontró un coeficiente de determinación de 0.92272778, el cual señala que el modelo da cuenta de 92.27% de la variación en y, lo cual era una señal de que se tenía un buen modelo.
2. En segundo lugar, se calculó la matriz de correlaciones y, con los criterios explicados antes, se volvió a construir un nuevo modelo de regresión, esta vez eliminando 3 variables independientes ("modelo de y y 4 variables i"), con lo que ya se tenía un modelo aceptable, en términos de las variables incluidas y en términos del nivel de explicación de la variación total.
3. En tercer lugar, se procedió a evaluar ese modelo reducido mediante el análisis de varianza y la prueba F. Aquí se realizaron pruebas sobre cada uno de los coeficientes de las variables independientes para evaluar si se podía concluir que eran diferentes de 0 y se concluyó que no se podía afirmar que el coeficiente de una de estas variables fuera diferente de 0, por lo que se le eliminó del análisis. Esto condujo a otro modelo con una variable menos: un "modelo de y y 3 variables i".

Al evaluar este último modelo se llegó a la conclusión de que se llegó al mejor modelo que fue posible construir con los datos disponibles.

4. A este "modelo de y y 3 variables i" se le aplicó el análisis de residuales para verificar que cumpliera en términos generales con los supuestos en los que se basa el modelo de regresión lineal múltiple y se concluyó que:

   a) El modelo cumplía aproximadamente con el supuesto de normalidad de los residuales.
   b) Sí se podía considerar que la relación entre las variables independientes con la dependiente era lineal.

Estos procedimientos recién referidos para los 2 ejemplos desarrollados detalladamente describen, en su conjunto y en forma más o menos completa, la regresión por pasos mediante eliminación posterior, ya que se comenzó incluyendo todas las variables disponibles y se eliminaron algunas en la medida que el proceso avanzaba, siguiendo los criterios mencionados, hasta llegar a un modelo satisfactorio.

14.7.2 Regresión por pasos mediante selección previa

En el caso de la regresión por pasos mediante selección previa, el procedimiento consiste básicamente en incorporar variables en el modelo, de una en una, hasta considerar todas. La forma en la que se suele utilizar es:

1. Se corre un modelo de regresión utilizando sólo 2 variables: la dependiente y la independiente que tenga la más alta correlación con aquélla.
2. De los pares de variables independientes que estén altamente correlacionadas entre sí, se excluyen del análisis, de cada par, la que tenga la menor correlación con la dependiente.
3. Se agrega otra variable independiente que esté altamente correlacionada con la dependiente pero que, a la vez, no lo esté con la independiente que ya se incorporó.

4. Se repiten los pasos 2 y 3, revisando en cada iteración si el cociente de varianza explicada a varianza total, es decir, el coeficiente de determinación múltiple, repasada en la sección 14.4.1, aumenta de manera significativa, lo cual querría decir que la introducción de cada variable nueva al modelo aumenta su capacidad de explicación porque, de no ser así, no tiene ningún valor introducir una variable adicional.

EJERCICIOS 14.7 Regresión por pasos

1. La tabla siguiente muestra los resultados de una investigación sobre 45 sujetos, en donde se obtuvieron mediciones sobre una variable dependiente y y 5 variables independientes x.

Sujeto	Y	x_1	x_2	x_3	x_4	x_5
1	79	16	38	71	18	27
2	68	14	56	80	13	29
3	76	19	64	92	22	48
4	79	15	38	73	22	29
5	71	14	40	71	26	36
6	72	12	28	61	23	20
7	59	11	39	65	19	41
8	73	15	48	79	18	35
9	63	16	48	56	24	41
10	76	15	38	82	23	48
11	63	14	58	74	22	32
12	67	12	31	62	19	38
13	83	18	61	97	11	39
14	91	17	27	75	13	48
15	64	13	48	104	16	38
16	83	16	38	79	19	36
17	72	15	44	78	20	40
18	78	17	46	74	18	41
19	51	10	48	61	23	32
20	82	14	30	86	20	52
21	65	14	48	59	12	29
22	83	19	51	88	28	39
23	72	14	42	72	15	34
24	70	14	48	74	8	43
25	73	13	28	59	11	44
26	84	15	31	80	16	48
27	46	11	49	45	15	31
28	91	15	19	70	8	44
29	63	11	42	70	22	25
30	75	15	37	70	21	44
31	67	13	50	78	17	30
32	95	16	36	103	10	42
33	88	16	41	100	13	40
34	70	14	50	80	16	31
35	68	15	53	79	17	50
36	97	19	38	87	9	46
37	80	18	59	89	16	35
38	86	17	45	88	20	39
39	73	12	38	74	12	27
40	81	15	29	70	12	50
41	71	15	41	71	26	36
42	63	16	49	56	24	41
43	91	17	28	75	14	49
44	79	18	47	74	18	41
45	63	12	42	70	22	25

Construya una ecuación de regresión lineal múltiple, por pasos, utilizando el método de selección previa y explique cada paso, junto con los resultados que se obtuvieron.

2. En un estudio sobre el efecto que tienen diversas variables sobre el costo de la mano de obra en el departamento de envíos de una empresa, se tomó una muestra de 20 meses y se recopiló información sobre la producción enviada y transportada: toneladas enviadas, porcentaje transportado por tren y número de paquetes por envío:

Mes	Horas de mano de obra	Toneladas enviadas	Porcentaje de unidades transportadas por tren	Núm. promedio de paquetes por envío
1	110	5.4	85	200
2	95	4.1	94	220
3	118	5.6	53	190
4	126	7.8	11	150
5	102	4.8	49	200
6	73	3.6	37	260
7	89	5.6	7	250
8	111	6.2	27	210
9	98	4.3	51	240
10	81	4.5	59	290
11	132	7.1	73	100
12	95	4.2	85	300
13	60	4.1	69	280
14	124	7.8	84	140
15	114	4.8	85	210

(*continúa*)

(*continuación*)

Mes	Horas de mano de obra	Toneladas enviadas	Porcentaje de unidades transportadas por tren	Núm. promedio de paquetes por envío
16	121	6.3	35	200
17	120	8.4	50	160
18	110	3.2	59	190
19	92	4.3	30	230
20	95	5.1	53	250

Construya una ecuación de regresión lineal múltiple, por pasos, utilizando el método de eliminación posterior y explique cada paso, junto con los resultados obtenidos.

3. En la tabla siguiente se presentan datos sobre precios y diversas características de departamentos nuevos en venta.

Precio	Superficie	Lugares de estacionamiento	Recámaras	Baños
1 690 000	78	1	2	1
2 185 000	130	1	2	2
2 165 000	130	1	3	2
2 250 000	143	1	3	2
2 299 000	169	1	3	1.7
2 350 000	169	2	3	2.5
2 399 000	169	1	3	2
2 479 000	221	2	3	2.5
2 600 000	247	2	3	2
2 699 000	234	1	3	2
2 349 000	169	1	4	2
2 550 000	234	1	4	2
2 699 000	221	2	4	3
2 945 000	260	2	4	3
3 099 000	273	2	4	3

Construya una ecuación de regresión lineal múltiple, por pasos, utilizando el método de selección previa y explique cada paso, junto con los resultados que se vayan obteniendo.

4. En la tabla siguiente se muestran datos de 18 acciones que cotizan en la Bolsa Mexicana de Valores, del sector de alimentos, bebidas y tabaco, para el 6 de octubre de 2011, según se publicaron al día siguiente en el periódico *El Financiero*:

Clave	Utilidad por acción al reporte	Utilidad por acción (últimos 12 meses)	Valor en libros	Múltiplo precio a utilidad al reporte	Rendimiento nominal (última semana)	Rendimiento nominal (último mes)	Rendimiento nominal (último año)
AGRIEXPA*	–0.9042	–0.9042	–0.6396	–0.07	–1.67	–1.67	–1.67
AC*	6.7881	2.94	13.59	8.35	–2.07	–2.07	–7.38
BACHOCO B	0.458	1.368	27.1939	53.93	–3.14	–3.14	–4.63
BAFAR B	0.274	0.724	8.9498	62.04	0	0	10.39
BIMBO A	25.7099	34.67	9.2937	1.1	7.05	7.05	–72.61
FEMSA UB	0.556	3.161	32.795	127.7	0	0	26.79
FEMSA UBD	0.556	3.161	32.795	162.21	0.03	0.03	32.3
GAM B	0.6	0.6	2.9534	11.92	0	0	297.22
GEUPEC B	–0.358	0.637	13.6267	–134.08	–3.81	–3.81	4.35
GMACMA B	–0.23	–0.136	1.9093	–4.35	0	0	0
GMODELO C	0.647	2.815	121.7194	122.36	–0.28	–0.28	4.9
GRUMA B	0.463	4.108	12.235	52.03	–1.19	–1.19	6.03
HERDEZ+	3.7705	2	7.8336	6.55	1.02	1.02	8.58
KOF L	1.54	5.133	263.1204	80.66	1.06	1.06	23.85
MASECA M	0.204	0.968	17.2883	67.16	–1.44	–1.44	–2
MINSA B	0.147	0.47	5.8585	67.11	0	0	17.65
NUTRISA+	0.664	2.095	8.9365	76.05	0	0	32.2
SAVIA A	–2.6249	–2.6249	1.3268	–0.3	0	0	0

Usando la utilidad por acción al reporte como variable dependiente, construya una ecuación de regresión lineal múltiple, por pasos, mediante el método de eliminación posterior y explicar cada paso, junto con los resultados obtenidos.

14.8 Resumen

En el análisis de regresión lineal múltiple se revisa la relación entre una variable y, dependiente, y 2 o más variables independientes $x_1, x_2, \ldots x_n$. El modelo de regresión lineal múltiple:

$$y = \beta_0 + \beta_1 x + \beta_2 x + \cdots \beta_n x + \varepsilon$$

se basa en los siguientes supuestos:

1. Es un modelo lineal en los parámetros.
2. Consta de una parte no aleatoria, la que se basa en las variables independientes, las x_i, y una parte aleatoria, los ε_i.
3. Para cada combinación de valores x_i existe una población de valores y_i que se distribuye en forma normal.
4. La varianza de todas la poblaciones de valores y_i que corresponden a cada combinación de valores x_i son todas iguales.
5. Los valores y_i son independientes entre sí.
6. Los valores ε_i son independientes entre sí.
7. Los términos del error, las ε_i se distribuyen de forma normal con media de 0.

Los modelos de regresión lineal múltiple, al igual que los de regresión simple, se construyen mediante el método de mínimos cuadrados y, en este capítulo, se usa Excel para realizar los cálculos.

Uno de los primeros pasos al construir un modelo de regresión lineal múltiple consiste en elaborar una tabla de correlaciones, con 2 propósitos: evitar incluir variables independientes que estén altamente correlacionadas entre sí, ya que esto podría crear problemas de multicolinealidad y, en segundo término, identificar las variables independientes que tienen la mayor correlación con la variable dependiente, ya que son éstas las que tienen el mejor potencial para contribuir a un modelo robusto que tenga un elevado coeficiente de determinación múltiple.

Ya que se tiene un modelo en principio satisfactorio, es necesario evaluarlo. Se llevan a cabo 3 tipos de evaluaciones.

1. Primero, se le evalúa en forma global, midiendo qué tan adecuada es la ecuación de regresión múltiple para medir la relación entre las variables. Esta evaluación se hace mediante: *a*) el coeficiente de determinación múltiple, y *b*) el análisis de varianza y la prueba con el estadístico F de Fisher.
2. Un segundo tipo de prueba sobre la ecuación de regresión es la que se hace sobre los coeficientes de regresión parciales individuales, con el propósito de evaluar si son estadísticamente significativos o, en otras palabras, para ver si son diferentes de 0 porque, en el caso de no poder probar que lo son, entonces las variables asociadas no son útiles en el modelo.
3. El tercer tipo de evaluación que se hace de la ecuación de regresión es la que se realiza sobre los residuales, la diferencia entre los valores observados de la variable dependiente y los valores estimados a través de la ecuación de regresión lineal múltiple que se construye. Esta evaluación de los residuales se relaciona, más que con la adecuación en sí del modelo, con el cumplimiento de los supuestos en los que se basa.

Una vez que se tiene un modelo adecuado, se le puede utilizar de 2 maneras: para construir intervalos de confianza para pronosticar los valores probables de la variable dependiente y estimar el promedio de la subpoblación de valores y que se asume existen para determinada combinación de las variables independientes.

Después de revisar estos usos de la ecuación de regresión lineal múltiple, se explicó el procedimiento que se puede seguir para construir modelos utilizando variables ficticias, las cuales se usan para integrar al análisis variables cualitativas.

Finalmente, se explicaron los procedimientos que se aplican para llevar a cabo lo que se conoce como *regresión por pasos* los cuales consisten en ir introduciendo o retirando variables independientes en el modelo de regresión, de 1 en 1, hasta analizarlas todas y llegar a un modelo satisfactorio.

14.9 Fórmulas del capítulo

El modelo de regresión lineal múltiple:

$$y = \beta_0 + \beta_1 x + \beta_2 x + \cdots \beta_n x + \varepsilon \qquad (14.1)$$

14.2 Obtención de la ecuación de regresión lineal múltiple

El conjunto de las ecuaciones normales para un modelo de regresión lineal múltiple con 2 variables independientes:

$$\sum y = nb_0 + b_1 \sum x_1 + b_2 \sum x_2 \qquad (14.2)$$

$$\sum x_1 y = b_0 \sum x_1 + b_1 \sum x_1^2 + b_2 \sum x_1 x_2 \qquad (14.3)$$

$$\sum x_2 y = b_0 \sum x_2 + b_1 \sum x_1 x_2 + b_2 \sum x_2^2 \qquad (14.4)$$

14.4.1 Evaluación de la ecuación de regresión mediante el coeficiente de determinación múltiple

El coeficiente de determinación múltiple ajustado:

$$r_{aj}^2 = 1 - \left[\left(1 - r^2\right) \frac{n-1}{n-k-1}\right] \qquad (14.5)$$

14.4.3 Inferencias sobre coeficientes de regresión parciales individuales

La t de Student como estadístico de prueba para los coeficientes individuales de la ecuación de regresión lineal múltiple:

$$t = \frac{b_i - \beta_i}{s_{b_i}} \qquad (14.6)$$

con $n - k - 1$ grados de libertad.

14.5.1 Intervalos de confianza para los pronósticos

Intervalo de confianza para pronósticos sobre valores probables de la variable dependiente:

$$\hat{y} \pm t_{\left(1-\frac{\alpha}{2},\, n-k-1\right)} s_{\hat{y}_i} \qquad (14.7)$$

14.5.1 Intervalos de confianza para estimaciones de la media de una subpoblación de valores y

Intervalo de confianza para estimaciones de la media de una subpoblación de valores y:

$$\hat{\mu}_y \pm t_{\left(1-\frac{\alpha}{2},\, n-k-1\right)} s_{\mu} \qquad (14.8)$$

14.10 Ejercicios adicionales

1. Explique la diferencia entre regresión lineal simple y múltiple.
2. Explique los siguientes conceptos:

   a) Matriz de correlación.
   b) Ecuación de regresión.
   c) Método de mínimos cuadrados.
   d) Coeficiente de determinación múltiple.
   e) Coeficiente de correlación múltiple.
   f) Multicolinealidad.
   g) Regresión por pasos mediante eliminación posterior.
   h) Regresión por pasos mediante selección previa.

14.3 Multicolinealidad y las variables que mejor se relacionan con la variable dependiente: uso de la matriz de correlaciones

3. ¿Por qué en la diagonal de las matrices de correlación el único valor es 1?
4. ¿Qué es la multicolinealidad?
5. ¿Por qué no se deben incluir en un modelo de regresión lineal múltiple 2 variables independientes que estén altamente correlacionadas entre sí?
6. En la siguiente matriz, la variable A es la variable dependiente:

	A	B	C	D
A	1			
B	−0.88	1		
C	0.79	−0.44	1	
D	0.23	−0.08	0.10	1

   a) ¿Conviene incluir en un modelo de regresión lineal múltiple a las variables B, C y D como variables independientes?
   b) ¿Habría problemas de multicolinealidad si se incluyen las 3 variables independientes en el modelo? Explique su respuesta.
   c) Indique cuáles variables se deben incluir en el modelo y por qué.

7. En una muestra de 10 supermercados se recopiló información sobre la proporción de clientes que preferían una marca de determinado producto, junto con el promedio de ingresos mensuales y el índice de escolaridad de las familias del área de influencia de cada supermercado, con los resultados que se muestran a continuación:

Supermercado	Proporción de preferencia de la marca	Promedio de ingresos mensuales	Índice de escolaridad
A	56.1	4	10.3
B	43.2	2.4	9.5
C	55.5	7.1	7.6
D	54.9	6.1	9.8
E	62.7	7.7	10.8
F	42.2	2.8	11.9
G	40.2	5.6	8.2
H	34	2.4	10
I	43.8	7.1	6.8
J	43.5	4.7	10.7

Calcule la matriz de correlaciones y considerando la proporción de consumidores que prefieren la marca en cuestión como la variable dependiente indique si ambas variables independientes contribuyen a un modelo de regresión lineal múltiple. Explique su respuesta.

8. Para evaluar la efectividad de los anuncios en periódicos y en radio para la promoción de cierto artículo, se recopilaron datos sobre cantidades de anuncios en ambos medios y en 20 ciudades de tamaño similar y se obtuvieron los siguientes resultados:

Ciudad	Ventas	Anuncios en periódicos	Anuncios en radio
1	1 070 300	440	0
2	1 230 900	440	0
3	962 500	275	275
4	687 500	275	275
5	1 001 000	330	330
6	1 068 100	330	330
7	1 024 100	385	385
8	1 294 700	385	385
9	970 200	275	440
10	1 080 200	275	440
11	1 790 800	495	495
12	1 734 700	495	495
13	1 148 400	0	550
14	1 005 400	0	550
15	1 461 900	275	605
16	1 463 000	275	605
17	1 545 500	330	660
18	1 579 600	330	660
19	1 673 100	385	715
20	1 915 100	385	715

Calcule la matriz de correlaciones y considerando las ventas como la variable dependiente indique si ambas variables independientes contribuyen a un modelo de regresión lineal múltiple. Explique su respuesta.

9. En pruebas de laboratorio se ensayaron diferentes combinaciones de componentes activos en un medicamento y se evaluó su efectividad. Los resultados que se obtuvieron fueron los siguientes:

Ingrediente *x*	Ingrediente *y*	Ingrediente *z*	Efectividad del medicamento
18	23	13	50
18	23	23	57
18	33	13	61
18	33	23	69
33	23	13	62
33	23	23	70
33	33	13	74
33	33	23	86
48	23	13	75
48	23	23	85
48	33	13	88
48	33	23	97

Calcule la matriz de correlaciones y considerando la efectividad del medicamento como la variable dependiente, indique si las 3 variables independientes contribuyen a un modelo de regresión lineal múltiple. Explique su respuesta.

10. En un estudio de hábitos de ahorro, se recopiló información sobre 15 familias:

Familia	Ahorro	Ingreso mensual	Habitación	Alimentos	Medicinas
1	660	8 250	2 805	3 135	825
2	0	6 270	2 475	2 970	330
3	1 650	13 860	4 620	3 960	1 320
4	1 320	11 220	3 960	4 290	0
5	495	9 240	3 630	3 300	990
6	825	10 230	2 640	4 125	825
7	2 640	17 160	4 950	4 620	1 815
8	0	14 520	5 775	5 115	1 485
9	3 135	11 880	2 970	2 805	660
10	990	12 705	3 465	4 455	1 155
11	165	6 765	2 640	3 465	0
12	495	8 745	2 145	3 135	495
13	660	6 435	1 650	2 640	330
14	0	8 250	2 970	3 300	825
15	1 485	15 840	4 620	5 280	1 485
16	1 634	17 424	5 082	5 808	1 634

Calcule la matriz de correlaciones y considerando el ahorro como la variable dependiente, indique si las 4 variables independientes contribuyen a un modelo de regresión lineal múltiple. Explique la respuesta.

14.4 Evaluación de la ecuación de regresión

11. Para los datos del ejercicio 7 sobre proporción de preferencia de una marca:

a) Determine el coeficiente de determinación múltiple e interprételo.

b) Plantee la hipótesis para realizar una prueba global de significación para la regresión utilizando el estadístico *F* de Fisher, interprételo y establezca la conclusión, tanto en términos de aceptar o rechazar la hipótesis nula como en términos de los datos mismos.

c) Si las conclusiones de *b*) muestran que tiene sentido, plantee hipótesis para realizar pruebas sobre los coeficientes de regresión parcial de la ecuación de regresión. Determine los estadísticos de prueba *t* de Student y establezca las conclusiones correspondientes.

d) ¿Las conclusiones de *c*) indican que se debe modificar el modelo? Si la respuesta es afirmativa, hágalo y repita los incisos *c*) y *d*).

e) Realice un análisis de los residuales del modelo al que se llegó hasta aquí:

i) ¿El histograma de los residuales tiene forma aproximadamente normal? Concluya con base en su respuesta.

ii) ¿Las cantidades de residuales positivos y negativos son aproximadamente iguales? Concluya con base en su respuesta.

iii) Analice el gráfico de probabilidad normal para esos residuales y fundamente la conclusión que se pueda extraer de esta gráfica.

12. Retome los datos del ejercicio 8 y realice lo siguiente:

a) Determine el coeficiente de determinación múltiple e interprételo.

b) Plantee la hipótesis para realizar una prueba global de significación para la regresión utilizando el estadístico *F* de Fisher, interprételo y establezca la conclusión, tanto en términos de aceptar o rechazar la hipótesis nula como en términos de los datos mismos.

c) Si las conclusiones de *b*) muestran que tiene sentido, plantee hipótesis para realizar pruebas sobre los coeficientes de regresión parcial de la ecuación de regresión. Determine los estadísticos de prueba *t* de Student y establezca las conclusiones correspondientes.

d) ¿Las conclusiones de *c*) indican que se debe modificar el modelo? Si la respuesta es afirmativa, hágalo y repita los incisos *c*) y *d*).

e) Realice un análisis de los residuales del modelo al que se llegó hasta aquí:

i) ¿El histograma de los residuales tiene forma aproximadamente normal? Concluya con base en su respuesta.

ii) ¿Las cantidades de residuales positivos y negativos son aproximadamente iguales? Concluya con base en su respuesta.

iii) Analice el gráfico de probabilidad normal para esos residuales y fundamente la conclusión que se pueda extraer de esta gráfica.

13. Para los datos del ejercicio 9, sobre diferentes combinaciones de componentes activos en un medicamento:

a) Determine el coeficiente de determinación múltiple e interprételo.

b) Plantee la hipótesis para realizar una prueba global de significación para la regresión utilizando el estadístico *F* de Fisher, interprételo y establezca la conclusión, tanto en términos de aceptar o rechazar la hipótesis nula como en términos de los datos mismos.

c) Si las conclusiones de *b*) muestran que tiene sentido, plantee hipótesis para realizar pruebas sobre los coeficientes de regresión parcial de la ecuación de regresión. Determine los estadísticos de prueba *t* de Student y establezca las conclusiones correspondientes.

d) ¿Las conclusiones de *c*) indican que se debe modificar el modelo? Si la respuesta es afirmativa, hágalo y repita los incisos *c*) y *d*).

e) Realice un análisis de los residuales del modelo al que se llegó hasta aquí:

i) ¿El histograma de los residuales tiene forma aproximadamente normal? Concluya con base en su respuesta.

ii) Las cantidades de residuales positivos y negativos son aproximadamente iguales? Concluya con base en su respuesta.

iii) Analice el gráfico de probabilidad normal para esos residuales y fundamente la conclusión que se pueda extraer de esta gráfica.

14. Para los datos del ejercicio 10, que trata de hábitos de ahorro:

a) Determine el coeficiente de determinación múltiple e interprételo.

b) Plantee la hipótesis para realizar una prueba global de significación para la regresión utilizando el estadístico *F* de Fisher, interprételo y establezca la conclusión, tanto en términos de aceptar o rechazar la hipótesis nula como en términos de los datos mismos.

c) Si las conclusiones de *b*) muestran que tiene sentido, plantee hipótesis para realizar pruebas sobre los coeficientes de regresión parcial de la ecuación de regresión. Determine los estadísticos de prueba *t* de Student y establezca las conclusiones correspondientes.

d) ¿Las conclusiones de *c*) indican que se debe modificar el modelo? Si la respuesta es afirmativa, hágalo y repita los incisos *c*) y *d*).

e) Realice un análisis de los residuales del modelo al que se haya llegado hasta aquí:

i) ¿El histograma de los residuales tiene forma aproximadamente normal? Concluya con base en su respuesta.

ii) ¿Las cantidades de residuales positivos y negativos son aproximadamente iguales? Concluya con base en su respuesta.

iii) Analice el gráfico de probabilidad normal para esos residuales y fundamente la conclusión que se pueda extraer de esta gráfica.

15. En la tabla siguiente se resumen los datos de 20 modelos de automóviles.

Auto	Tamaño del motor (cm^3)	Peso (kg)	Longitud	Rendimiento de gasolina (km/l)
1	3 020	2 148	4.50	5.67
2	1 400	1 225	3.23	10.00
3	4 400	2 500	4.78	4.00
4	3 510	2 263	4.60	5.00
5	3 500	2 035	4.50	6.00
6	3 050	1 925	4.30	6.00
7	850	1 013	2.98	11.00
8	3 050	1 983	4.25	5.67
9	1 300	1 495	3.83	7.33
10	2 320	1 600	3.25	6.67
11	3 180	2 070	4.45	5.67
12	850	1 010	3.08	9.67
13	850	985	2.93	11.33
14	4 000	2 343	4.68	4.33
15	2 500	1 675	3.93	7.33
16	2 250	1 685	4.03	7.67
17	910	880	2.75	14.67
18	2 500	1 688	3.95	8.00
19	970	1 133	3.13	10.67
20	1 110	1 078	3.28	9.00

Ajuste una ecuación de regresión lineal múltiple, con el rendimiento de la gasolina como variable dependiente, y:

a) Determine el coeficiente de determinación múltiple e interprételo.

b) Plantee la hipótesis para realizar una prueba global de significación para la regresión utilizando el estadístico *F* de Fisher, interprételo y establezca la conclusión, tanto en términos de aceptar o rechazar la hipótesis nula como en términos de los datos mismos.

c) Si las conclusiones de *b*) muestran que tiene sentido, plantee hipótesis para realizar pruebas sobre los coeficientes de regresión parcial de la ecuación de regresión. Determine los estadísticos de prueba *t* de Student y establezca las conclusiones correspondientes.

d) ¿Las conclusiones de *c*) indican que se debe modificar el modelo? Si la respuesta es afirmativa, hágalo y repita los incisos *c*) y *d*).

e) Realice un análisis de los residuales del modelo al que se haya llegado hasta aquí:

i) ¿El histograma de los residuales tiene forma aproximadamente normal? Concluya con base en su respuesta.

ii) ¿Las cantidades de residuales positivos y negativos son aproximadamente iguales? Concluya con base en su respuesta.

iii) Analice el gráfico de probabilidad normal para esos residuales y fundamente la conclusión que se pueda extraer de esta gráfica.

16. Un despacho de auditores analiza la relación entre el tiempo que sus auditores dedican a su labor, junto con el tiempo de computadora que se utiliza y la magnitud de los errores contables que se detectan y encuentra los siguientes datos para los meses de un año determinado:

Mes	Magnitud de los errores contables	Horas de trabajo de los auditores	Horas de computadora
Ene.	500 000	21	34
Feb.	470 000	19	29
Mar.	490 000	20	32
Abr.	500 000	18	30
Mayo	480 000	18	31
Jun.	510 000	19	33
Jul.	490 000	21	35
Ago.	500 000	21	33
Sept.	490 000	20	33
Oct.	480 000	20	32
Nov.	520 000	21	29
Dic.	550 000	22	31

Ajuste una ecuación de regresión lineal múltiple, con la magnitud de los errores contables como variable dependiente, y:

a) Determine el coeficiente de determinación múltiple e interprételo.

b) Plantee la hipótesis para realizar una prueba global de significación para la regresión utilizando el estadístico F de Fisher, interprete éste y establezca la conclusión, tanto en términos de aceptar o rechazar la hipótesis nula como en términos de los datos mismos.

c) Si las conclusiones de *b*) muestran que tiene sentido, plantee hipótesis para realizar pruebas sobre los coeficientes de regresión parcial de la ecuación de regresión. Determine los estadísticos de prueba t de Student y establezca las conclusiones correspondientes.

d) ¿Las conclusiones de *c*) indican que se debe modificar el modelo? Si la respuesta es afirmativa, hágalo y repita los incisos *c*) y *d*).

e) Realice un análisis de los residuales del modelo al que se llegó hasta aquí:

i) ¿El histograma de los residuales tiene forma aproximadamente normal? Concluya con base en su respuesta.

ii) ¿Las cantidades de residuales positivos y negativos son aproximadamente iguales? Concluya con base en su respuesta.

iii) Analice el gráfico de probabilidad normal para esos residuales y fundamente la conclusión que se pueda extraer de esta gráfica.

14.5 Uso del modelo de regresión lineal múltiple

17. Con los resultados de los ejercicios 7 y 11, construya un intervalo de confianza de 95% para pronosticar el valor de $\hat{y}$, utilizando los siguientes valores para las variables independientes, en caso de que quedar incluidas en el modelo final.

Promedio de ingresos mensuales	5.8
Índice de escolaridad	10

18. Con los resultados de los ejercicios 8 y 12 construya un intervalo de confianza de 99% para pronosticar el valor de $\hat{y}$, utilizando los siguientes valores para las variables independientes, en caso de que quedaran incluidas en el modelo final.

Anuncios en periódicos	400
Anuncios en radio	400

19. Con los resultados de los ejercicios 9 y 13, construya un intervalo de confianza de 90% para pronosticar el valor de $\hat{y}$, utilizando los siguientes valores para las variables independientes, en caso de que quedaran incluidas en el modelo final.

Ingrediente *x*	20
Ingrediente *y*	20
Ingrediente *z*	20

20. Con los resultados de los ejercicios 10 y 14, construya un intervalo de confianza de 95% para pronosticar el valor de $\hat{y}$, utilizando los siguientes valores para las variables independientes, en caso de que quedaran incluidas en el modelo final.

Ingreso mensual	10 000
Habitación	4 000
Alimentos	4 000
Medicinas	1 000

21. Con los resultados del ejercicio 15, construya un intervalo de confianza de 98% para pronosticar el valor de $\hat{y}$, utilizando los siguientes valores para las variables independientes, en caso de que quedaran incluidas en el modelo final.

Tamaño del motor	2 500
Peso	2 000
Longitud	4

22. Con los resultados del ejercicio 16, construya un intervalo de confianza de 96% para pronosticar el valor de $\hat{y}$, utilizando los siguientes valores para las variables independientes, en caso de que quedaran incluidas en el modelo final.

Horas de trabajo de los auditores	22
Horas de computadora	40

14.6 Variables independientes cualitativas

23. Para evaluar la relación que existe entre la edad y la hipertensión con el estado general de salud, el cual resume la calificación sobre 10 indicadores, se tomó una muestra de 23 sujetos y se obtuvieron los siguientes resultados, en donde

1 representa un individuo con hipertensión y 0 otro que no la padece.

Estado de salud	Edad	Hipertenso
69.9	33	1
73	37	0
53.9	57	0
54.7	45	1
40.1	74	0
61.8	33	1
70.9	43	0
40	63	1
49	73	0
67.5	39	1
65	55	0
48.2	58	1
77.9	36	0
69.7	47	0
54.2	38	1
48.2	68	0
47.1	50	1
84	42	0
56.1	62	0
51.4	40	1
70.2	49	0
35.9	58	1
46.4	51	1

a) Construya una ecuación de regresión lineal múltiple que describa la relación del estado de salud como variable dependiente y plantee las conclusiones a las que se pueda llegar a partir del modelo.
b) Interprete el coeficiente de la variable ficticia.

24. Para evaluar el efecto de la capacitación y el turno laboral sobre la calidad de la producción, se tomó una muestra de 40 turnos de trabajo en la fábrica y se determinó la cantidad de horas que los operarios dedicaron a capacitación, junto con la cantidad de artículos sin defectos que lograron elaborar en cada uno de los turnos matutino (I), vespertino (II) y nocturno (III).

Artículos sin defectos	Horas de capacitación	Turno
2 050	410	I
2 360	520	I
2 600	750	II
1 760	610	III
2 450	800	I
1 230	340	I
1 760	600	II
1 500	380	III
1 480	300	III
2 650	800	II
2 000	480	I
450	250	II
1 100	250	III
2 160	750	II
1 760	370	I
900	250	III
1 760	310	I
1 120	350	III
2 300	700	II
2 800	770	I
2 200	780	II
2 600	700	I
450	310	II
2 100	700	III
2 250	600	I
1 260	550	II
2 000	740	III
1 150	290	I
650	310	II
2 400	750	I
1 940	670	II
1 560	520	III
1 000	250	I
2 400	790	II
1 700	440	I
1 160	420	III
1 200	500	II
2 400	630	I
880	440	II

a) Construya una ecuación de regresión lineal múltiple que describa la relación del número de artículos sin defectos como variable dependiente y plantee las conclusiones a las que se pudiera llegar a partir del modelo.
b) Interprete los coeficientes de las variables ficticias.

25. Construya un modelo de regresión lineal múltiple para los siguientes datos sobre casas utilizando el precio como variable dependiente e interprete los coeficientes de las variables ficticias.

Casa	Precio	Metros cuadrados de construcción	Metros cuadrados de terreno	Ubicación
1	$3 110 000	125	400	A
2	$2 697 500	100	333	A
3	$2 395 000	100	333	A
4	$3 500 000	133	500	A

Casa	Precio	Metros cuadrados de construcción	Metros cuadrados de terreno	Ubicación
5	$2 850 000	117	400	A
6	$2 785 000	100	333	A
7	$3 680 000	133	500	A
8	$3 480 000	125	400	A
9	$3 540 000	125	400	A
10	$2 530 000	108	400	A
11	$3 450 000	133	433	B
12	$4 650 000	158	500	B
13	$4 700 000	150	500	B
14	$4 015 000	167	567	B
15	$4 850 000	167	567	B
16	$4 120 000	142	500	B
17	$4 085 000	150	500	B
18	$4 300 000	158	533	B
19	$4 400 000	167	533	B
20	$3 875 000	133	433	B
21	$5 935 000	200	600	C
22	$7 340 000	217	600	C
23	$6 620 000	192	533	C
24	$5 560 000	183	533	C
25	$6 955 000	233	667	C

26. Para los datos que aparecen en la tabla siguiente, sobre el sexo, la antigüedad y el sueldo de 24 maestros de bachillerato, construya una ecuación de regresión lineal múltiple utilizando el sueldo como variable dependiente, comente sobre la adecuación del modelo e interprete el coeficiente de la variable ficticia.

Maestro	Sexo	Antigüedad	Sueldo
1	H	11	16
2	M	9	15.2
3	M	8	15
4	H	13	18.1
5	M	16	19.9
6	H	22	22
7	H	23	20.1
8	M	11	17.5
9	M	9	16.2
10	M	20	20.3
11	M	8	11.5
12	M	13	16.5
13	H	16	16.7
14	M	13	16
15	H	18	26.5
16	M	10	17
17	M	8	18.1
18	H	13	18.1
19	M	9	12
20	H	11	12.5
21	M	15	13.2
22	M	4	11
23	M	13	18
24	H	11	18

27. Con los datos de la tabla siguiente, sobre 15 ejecutivos, construya una ecuación de regresión lineal múltiple utilizando el sueldo como variable dependiente, comente sobre la adecuación del modelo e interprete el coeficiente de la variable ficticia.

Ejecutivo	Sueldo mensual	Años de experiencia	Años de escolaridad	Sexo
1	23 023	7.5	13	F
2	27 335	11	13	M
3	26 103	6	14	F
4	26 180	10	13	M
5	25 025	11.5	14	M
6	23 485	5	13	F
7	23 870	9	12	F
8	21 329	3.5	13.5	F
9	30 800	10.5	14	M
10	26 950	9.5	15	F
11	23 870	11.5	11	M
12	22 022	8	11	F
13	23 100	4.5	13	M
14	21 175	3.5	13.5	M
15	23 870	7	14	F

14.7 Regresión por pasos

28. Utilizando los datos siguientes sobre diversas características de zonas urbanas similares en 25 ciudades, y el porcentaje de viviendas y el de viviendas desocupadas como variables dependientes, construya una ecuación de regresión lineal múltiple, por pasos, con el método de selección previa y explique cada paso, junto con los resultados obtenidos.

Ciudad	Promedio del ingreso familiar	Promedios de los sueldos	Promedio de los alquileres de vivienda	% de desempleo o subempleo	% de viviendas desocupadas
1	9 647	6 700	4 660	18.3	6.2
2	9 309.5	4 844	5 020	16.23	6.7
3	9 725.5	8 630.5	5 260	15.7	7
4	10 071	8 701	4 420	2.84	2.9
5	10 352	10 665.5	4 320	9	2.9
6	9 134.5	6 553.5	5 020	33.56	8.6

(*continúa*)

(*continuación*)

Ciudad	Promedio del ingreso familiar	Promedios de los sueldos	Promedio de los alquileres de vivienda	% de desempleo o subempleo	% de viviendas desocupadas
7	11 561.5	10 491.5	4 920	–0.04	2.8
8	9 326	8 454.5	4 780	10.16	7.1
9	10 257	9 802.5	4 620	4.43	3.1
10	11 777	6 651.5	6 300	64.89	16.5
11	10 386.5	8 890	5 200	0.87	3
12	10 932	8 803.5	4 440	6.98	3.2
13	10 875.5	9 902.5	4 560	12.43	3.9
14	11 193	10 137.5	5 040	3.67	3.1
15	9 406.5	9 194	3 860	2.46	4.8
16	9 693	10 403.5	4 640	32.61	7.9
17	10 895.5	9 387	4 880	25.1	4
18	10 394.5	11 250	4 460	9.26	3.2
19	12 259.5	9 169.5	5 040	2.95	3
20	9 485.5	7 871.5	4 640	–0.89	5
21	9 064.5	7 073	4 480	39.78	8.3
22	7 642	6 752.5	3 300	0.33	3.8
23	10 877	9 593	4 520	16.17	5
24	8 699.5	7 385	4 060	0.6	3.5
25	10 372	9 129.5	4 620	2.16	3.2

29. En la tabla siguiente se muestran datos de 25 acciones que cotizan en la Bolsa de Valores de Nueva York.

Acción	Precio actual	Cambio en ventas	Cambio en ingresos	Cambio en activos	Dividendo anual	Utilidad por acción	Precio de cierre un año antes
1	51.35	4.56	–16.06	3.84	1.584	1.989	38.5
2	61.75	1.68	76.67	13.68	1.152	3.348	33
3	47.45	–18.84	97.57	0.72	0.9	2.232	31.9
4	42.9	3.6	4.4	3.12	1.422	2.124	29.7
5	118.625	12.6	24.64	31.32	0.252	1.647	49.225
6	9.1	14.16	–14.52	29.88	0.099	0.333	20.075
7	71.5	3	17.27	7.56	1.17	3.591	41.525
8	44.044	16.08	7.59	77.64	0.36	0.936	24.618
9	24.05	20.52	42.13	23.64	0.225	0.9	18.975
10	44.369	–3.96	29.04	8.88	0.819	2.475	24.2
11	30.719	3.12	4.18	5.28	0.828	1.125	23.925
12	39.65	29.4	104.17	65.04	0.072	0.927	15.95
13	28.6	14.04	78.98	10.44	0.819	1.26	18.425
14	28.444	7.2	52.8	27.96	0.45	1.557	15.268
15	68.9	–9	85.14	–1.2	0.477	4.905	30.25
16	27.794	7.44	16.39	10.56	0.351	1.17	16.775
17	17.069	32.04	–13.86	19.68	0.18	0.549	10.043
18	31.044	–7.92	200.53	12.12	0.567	1.494	14.025
19	33.969	19.56	61.82	10.92	0.513	1.557	24.475
20	52.65	–10.8	19.8	8.52	1.026	2.79	31.493
21	29.25	–11.52	23.87	8.52	0.594	1.809	17.468
22	27.3	36.48	47.3	62.04	0	0.765	30.525
23	22.425	29.28	64.79	9.72	0.675	0.981	13.068
24	26.325	50.88	63.8	30.36	0.054	0.81	11.693
25	34.619	22.68	42.13	27.6	0.882	1.611	18.018

Construya una ecuación de regresión lineal múltiple, por pasos, con el método de eliminación posterior y el precio actual como variable dependiente, y explicar cada paso, junto con los resultados obtenidos.

30. Con los datos siguientes del desempeño de una cadena de tiendas de conveniencia, construya un modelo de regresión lineal múltiple, utilizando el promedio mensual de ventas como variable dependiente, siguiendo el método por pasos de selección previa y documentando los resultados obtenidos en cada paso.

Año	Núm. de tiendas	Tamaño promedio de las tiendas (m^2)	Promedio mensual de ventas (miles de pesos)	Núm. de empleados
1	15.5	616	519.6	2 000
2	25	640	841.2	2 700
3	30	640	1 201.2	3 300
4	37.5	656	1 744.8	4 550
5	48	688	2 400	6 500

Año	Núm. de tiendas	Tamaño promedio de las tiendas (m^2)	Promedio mensual de ventas (miles de pesos)	Núm. de empleados
6	59	704	3 309.6	8 750
7	72.5	736	4 578	10 750
8	87	760	6 163.2	14 000
9	107	784	8 577.6	19 450
10	132	800	11 086.8	25 300

capítulo 15

Números índice

Sumario

15.1 Usos de los números índice
15.2 Números índice simples
15.3 Números índice agregados
15.4 Números índice agregados de Laspeyres, de Paasche e ideal de Fischer
15.4.1 Índice de Laspeyres
15.4.2 Índice de Paasche
15.4.3 Índice ideal de Fischer
15.5 Números índices en cadena
15.5.1 Números índice en cadena y rendimientos bursátiles
15.6 Índices para propósitos especiales
15.6.1 Índices de precios al consumidor y al productor
15.6.2 Índices bursátiles
15.7 Números índices y Excel
15.8 Resumen
15.9 Fórmulas del capítulo
15.10 Ejercicios adicionales

Número índice. Valor relativo expresado en forma de porcentaje; mide precios, cantidades y valores durante un periodo dado contra el correspondiente precio en un periodo base.

Un **número índice** es un valor relativo expresado en forma de porcentaje y mide, por ejemplo, el precio de algún artículo en un periodo dado contra el correspondiente precio en un periodo base: si el precio de un artículo en el año 1, el año base, fue de $50 y el precio al año siguiente, el año 2, fue de $55,

$$IP = \frac{P_1}{P_0}(100) = \frac{55}{50}(100) = 110$$

Así, 110 es el índice de precio para ese artículo y es fácil ver que el 10 muestra que dicho precio subió 10% en un año, en tanto que el 100 restante representa el precio del artículo en el año base. Para apreciar esto de mejor manera se puede dividir la operación anterior de la siguiente forma:

$$IP = \frac{P_1}{P_0}(100) = \frac{P_0}{P_0} + \frac{P_1}{P_0} = \left(\frac{50}{50} + \frac{5}{50}\right)(100) = (1+0.1)(100) = 110$$

De la misma manera que se construye este sencillo índice de precios se crean también índices de cantidad e índices de valor en donde éstos se obtienen combinando (multiplicando) los precios y las cantidades de los artículos.

Índice simple. Es el índice calculado con los datos de un solo artículo en valor, precio o cantidad.
Índice agregado. Es el índice que se calcula con los datos de varios artículos en valor, precio o cantidad.

Los índices calculados como se acaba de ilustrar son **índices simples** porque se calculan con los datos de un solo artículo. Se revisan ejemplos adicionales de números índice simples en la sección 15.2. Cuando el índice se calcula con los datos de varios artículos, entonces se tiene un **índice agregado** y también se pueden calcular este tipo de índices para precios, cantidades y valores. Se revisa este tipo en la sección 15.3.

Existen además algunos números índice agregados especialmente importantes, creados por estadísticos destacados, como los de Fischer, Paasche y Laspeyres, los cuales se revisan en la sección 15.4.

Finalmente, existen también diversos números índice para propósitos especiales y que son muy conocidos. Los principales son los índices de precios al consumidor y al productor, y los índices bursátiles como el *Índice de Precios y Cotizaciones del mercado accionario* de la Bolsa Mexicana de Valores, o el Índice Dow Jones de la Bolsa de Valores de Nueva York (New York Stock Exchange), los cuales se analizan en la sección 15.5.

Para completar esta introducción a los números índice, en la sección siguiente se hace una breve exposición de las formas en las que pueden ser útiles.

15.1 Usos de los números índice

Los 3 usos principales de los números índice son los siguientes:

1. En primer lugar, permiten comparar los cambios en series disímbolas de datos. Siendo, como se acaba de exponer, valores relativos (expresados en porcentajes) que permiten comparar, por ejemplo, datos de precios contra datos de producción; los datos de precios podrían darse en decenas en tanto que los datos de producción podrían estar en miles de toneladas. Otro ejemplo de lo anterior tiene que ver con el hecho mencionado de que se trata de números relativos, al expresar las cantidades originales como porcentaje ya no se habla, por ejemplo, de un aumento de 1 245 587 toneladas de frijol sino de un aumento de 25% respecto al año anterior; es claro que la expresión porcentual es mucho más accesible que la original en toneladas.
2. La segunda utilidad que se desea destacar se refiere a los números índice agregados pues, al utilizarlos, se pueden comparar cambios en conjuntos de artículos. Los índices de precios al consumidor y al productor son el ejemplo clásico de esto ya que, como se verá, se construyen a partir de los precios y cantidades de las canastas básicas de artículos.
3. Finalmente, los números índice se pueden utilizar para eliminar en series de datos los efectos de la inflación, proceso conocido como *deflacionar*, el cual conduce a estimaciones de ingresos o gastos reales o a estimaciones del poder de compra real de una moneda. En el mismo sentido, se pueden utilizar también para eliminar los efectos de otro tipo de influencias como los movimientos estacionales en los precios, como los que se producen en tiempos vacacionales o navideños, lo que permite apreciar de mejor manera el comportamiento de los precios a más largo plazo. Se ve esto en el capítulo 16, que se ocupa del análisis de series de tiempo.

15.2 Números índice simples

Tal como ya se anotó, los números índice simples se calculan sobre un solo artículo.

Ejemplo 15.1

En la siguiente tabla se muestran los precios y los volúmenes de producción en México de 3 productos agrícolas para los años de 1998 a 2008 con datos del Instituto Nacional de Estadística y Geografía (INEGI).

Tabla 15.1 Precios y volúmenes de producción en México de 3 productos agrícolas, 1998–2008.

	Precio			Volumen (miles de toneladas)		
Año	Ajonjolí	Arroz palay	Cártamo	Ajonjolí	Arroz palay	Cártamo
1998	4.94	1.64	2.18	31.65	458.11	171.22
1999	5.73	1.78	1.95	31.46	326.51	262.74
2000	5.66	1.47	1.61	40.78	351.45	96.44
2001	4.96	1.48	1.31	42.88	226.64	111.46
2002	5.17	1.64	1.78	20.21	227.19	52.86
2003	6.25	1.66	2.27	31.03	273.27	200.59
2004	7.68	1.82	2.35	33.09	278.54	230.87
2005	7.95	1.90	2.26	20.04	291.15	94.42
2006	6.99	1.91	2.34	21.26	337.25	73.54
2007	8.00	2.08	2.36	29.05	294.70	113.33
2008	9.68	3.63	3.71	34.32	224.37	95.83

Calcule los índices simples de precios, cantidades y valores para los 3 productos, para 2008, utilizando como año base 1998.

Solución: Los índices de precios son:

Ajonjolí:

$$IP = \frac{9.68}{4.94}100 = 196.95$$

Arroz:

$$IP = \frac{3.63}{1.64}100 = 221.34$$

Cártamo:

$$IP = \frac{3.71}{2.18}100 = 170.18$$

Los índices de cantidad (de producción) son:

Ajonjolí:

$$IQ = \frac{34.32}{31.65}100 = 108.44$$

Arroz:

$$IQ = \frac{224.37}{458.11}100 = 48.98$$

Cártamo:

$$IQ = \frac{95.83}{171.22}100 = 55.97$$

Los índices de valor son:

Ajonjolí:

$$IV = \frac{9.68(34.32)}{4.94(31.65)}100 = 212.48$$

Arroz:

$$IV = \frac{3.63(224.37)}{1.64(458.11)}100 = 108.41$$

Cártamo:

$$IV = \frac{3.71(95.83)}{2.18(171.22)}100 = 95.25$$

Estos resultados permiten obtener información importante. En primer lugar, los índices de precios indican que el arroz aumentó de precio en más del doble, en tanto que los otros 2 productos casi alcanzaron el doble. En términos de la producción, los índices muestran que la de ajonjolí creció apenas 8.44%, en tanto que las de arroz y de cártamo se redujeron a cerca de la mitad.

Por su parte, los índices de valor que, como se vio, se obtienen combinando (multiplicando) los precios y los volúmenes de producción, indican que los valores de la producción de arroz y de cártamo se mantuvieron aproximadamente iguales (la de arroz subió 8.41% y la de cártamo bajó 4.75%), el valor de la producción de ajonjolí subió un poco más del doble: 112.48 por ciento.

Como puede verse en este ejemplo, aunque los índices simples son, efectivamente, fáciles de calcular, ofrecen información útil. Observando las operaciones anteriores es fácil ver que las fórmulas para calcular estos índices simples son:

$$ISP = \frac{P_n}{P_0}(100) \tag{15.1}$$

En donde, P_n es el precio del artículo en el tiempo (periodo) de interés, 2008 en el ejemplo, y P_0 es el precio en el periodo base, 1998, en el ejemplo. La multiplicación por 100 se debe a que los números índice se plantean en porcentaje, aunque no se les suele incluir el conocido símbolo %. Además, en este caso se anotaría 1998 = 100, para denotar que el periodo base es el año 1998.

Siguiendo la idea

$$ISQ = \frac{Q_n}{Q_0}(100) \tag{15.2}$$

En donde Q representa, por supuesto, cantidad.
La fórmula correspondiente al índice de valor es:

$$ISV = \frac{P_n Q_n}{P_0 Q_0}(100) \tag{15.3}$$

Es importante incluir aquí una nota respecto a este índice de valor, ya que puede propiciar confusiones si se consulta a autores diversos que abordan este tema de los números índice. La idea de la media o promedio ponderado que se vio en el capítulo 3, en la sección 3.1.2 para ser precisos, en donde se explica que la media ponderada se utiliza para darle un peso relativo diferente a cada uno de los valores; en el caso del índice de valor se puede considerar que se trata de un índice de precios ponderado por las cantidades o, recíprocamente, un índice de cantidad ponderado con los precios. Conviene tener esto presente porque algunos autores denominan a estos índices de valor, *índices ponderados* pero, por otra parte, como es claro que la multiplicación del precio de un artículo por la cantidad produce, precisamente, su valor es conveniente visualizar estos índices como índices de valor.

ejercicios 15.2 Números índice simples

En la tabla 15.2 se muestran los precios de diferentes artículos a lo largo de varios años durante el mes de diciembre, así como el consumo mensual de cada uno de ellos.

Tabla 15.2 Precios y consumos mensuales de diversos productos

Producto	Unidad	Precio promedio				Consumo *per cápita* mensual			
		2002	2003	2004	2005	2002	2003	2004	2005
Aceite	L	8.40	9.40	11.80	12.22	1.2	0.87	1.17	1.28
Jabón de tocador	Barra	5.58	5.60	5.63	5.50	0.7	0.73	0.83	0.85
Pañales	Paquete	54.80	54.80	55.39	62.01	8	8.7	9.3	11.5
Detergente en polvo para trastes	Kg	13.50	14.11	13.72	14.72	0.73	0.85	0.91	1.05
Azúcar estándar morena	Bolsa (2 kg)	13.64	14.70	17.90	22.30	0.35	0.38	0.48	0.52
Refresco de cola	Lata (355 ml)	4.10	4.20	4.13	5.07	12	18	23	27
Pechuga de pollo	Kg	31.00	28.70	33.35	34.10	17	22	26	31
Lenteja	Bolsa (500 g)	4.80	6.45	6.46	7.30	0.41	0.44	0.53	0.61
Mantequilla	Barra (225 g)	7.75	7.40	8.55	10.99	0.83	1.04	1.3	1.8
Queso panela	Kg	52.52	48.52	57.68	62.89	0.09	0.12	0.17	0.23
Arroz	Kg	3.40	4.40	4.83	6.90	0.86	1.2	1.33	1.42
Harina de trigo	Kg	5.20	4.65	5.70	6.05	0.11	0.18	0.23	0.27
Bolillo	Pieza	0.78	0.80	0.80	0.90	24	27	31	34
Pasta para sopa	Paquete (200 g)	1.85	1.95	1.94	1.95	16	19	24	28
Atún en aceite	Lata (174 g)	—	5.15	6.14	7.90	—	4	5.3	6.02

1. Calcule el índice simple de precio de la pechuga de pollo, lenteja y mantequilla para el año 2005, con 2002 = 100.
2. Calcule el índice simple de precio del refresco de cola para 2003, 2004 y 2005 tomando como base 2002.
3. Calcule el índice simple de precio del detergente en polvo para trastes para el año 2005, con 2002 = 100, y para el 2004 tomando como base 2003.
4. Calcule el índice simple de precio del azúcar para 2003, 2004 y 2005 tomando como base 2002.
5. Calcule el índice simple de cantidad del aceite para todos los años, con 2002 = 100.
6. Calcule el índice simple de cantidad del jabón de tocador y queso panela para el año 2005 tomando como base 2003.
7. Calcule el índice simple de cantidad del arroz, harina de trigo y pasta de sopa para 2004, con 2002 = 100.
8. Calcule el índice simple de valor de los pañales para 2003, 2004 y 2005, tomando como base 2002.
9. Calcule el índice simple de valor del bolillo y atún para 2003, 2004 y 2005, con 2002 = 100.
10. Calcule el índice simple de valor del refresco de cola y aceite para 2005 tomando como base 2003.

15.3 Números índice agregados

Muchas veces es necesario conocer el cambio en los precios de más de un producto; para medir el cambio combinado de los precios en un grupo de artículos se calcula el índice de precios agregados.

Para obtener un índice de precios agregados simplemente se suman los precios de los productos de interés, tanto para el periodo base, el periodo 0, como para el periodo de interés, el periodo n, y se dividen las sumas para, finalmente, multiplicarlas por 100. Se sigue el mismo procedimiento para calcular índices de cantidad y de valor.

Las fórmulas que resumen los procedimientos descritos son:

$$IAP = \frac{\sum P_n}{\sum P_0}(100) \tag{15.4}$$

$$IAQ = \frac{\sum Q_n}{\sum Q_0}(100) \tag{15.5}$$

$$IAV = \frac{\sum P_n Q_n}{\sum P_0 Q_0}(100) \tag{15.6}$$

EJEMPLO 15.2

En la siguiente tabla se muestran los precios de los principales productos necesarios para el mantenimiento de un automóvil para 2007 y 2010, junto con las cantidades que un usuario compró en esos 2 años. Calcule el índice de precios agregados, con 2007 = 100.

Artículo	2007	2010
Gasolina (L)	7.41	10
Botella aceite (480 ml)	128	180
Llantas	685	1 300
Anticongelante (250 ml)	40	60

Solución: El índice de precios es, siguiendo la fórmula 15.4:

$$IAP = \frac{\sum P_n}{\sum P_0}(100) = \frac{10 + 180 + 1\,300 + 60}{7.41 + 128 + 685 + 40}(100)$$

$$= \frac{1\,550}{860.41}(100) = 180.$$

Este número índice indica que los precios de esos 4 productos aumentaron 80% en conjunto, de 2007 a 2010.

EJEMPLO 15.3

Utilizando los datos de la tabla 15.1 construya índices de precios, de cantidades y de valor, para los 3 productos incluidos: ajonjolí, arroz palay y cártamo para 2008, utilizando 1998 como base (1998 = 100).

Solución: Utilizando la fórmula (15.4), el índice agregado de precios 2008, con 1998 = 100, es:

$$IAP = \frac{\sum P_n}{\sum P_0}(100) = \frac{9.68 + 3.63 + 3.71}{4.94 + 1.64 + 2.18}(100)$$

$$= \frac{17.02}{8.76}(100) = 194.29$$

Con la fórmula (15.5), el índice agregado de cantidades 2008 y 1998 = 100, es:

$$IAQ = \frac{\sum Q_n}{\sum Q_0}(100) = \frac{34.32 + 224.37 + 95.83}{31.65 + 458.11 + 171.22}(100)$$

$$= \frac{354.52}{660.98}(100) = 53.64$$

En donde este índice señala que las cantidades producidas de esos 3 artículos bajaron en conjunto a aproximadamente la mitad.

Con la fórmula (15.6), el índice agregado de valor 2008 y 1998 = 100, es:

$$IAV = \frac{\sum P_n Q_n}{\sum P_0 Q_0}(100)$$

$$= \frac{9.68(34.32) + 3.63(224.37) + 3.71(95.83)}{4.94(31.65) + 1.64(458.11) + 2.18(171.22)}(100)$$

$$= \frac{1\,502.21}{1\,280.91}(100) = 117.27$$

Este índice indica que el valor de la producción de esos 3 productos aumentó 17.27%, de 1998 a 2008.

15.4 Números índice agregados de Laspeyres, de Paasche e ideal de Fischer

Tal como se mencionó en la introducción, existen diversos índices de valor especialmente diseñados por estadísticos destacados. Los 3 principales de ellos son los que se conocen como índice de Laspeyres, índice de Paasche y el índice ideal de Fischer.

Índice de Laspeyres. Se distingue porque utiliza las mismas cantidades del año base, para calcular los índices agregados de valor de ponderación fija.

15.4.1 Índice de Laspeyres

Este índice se distingue porque utiliza las mismas cantidades del año base, para calcular los índices agregados de valor de ponderación fija.

Se vio antes que los números índice agregados de valor se calculan como:

$$IAV = \frac{\sum P_n Q_n}{\sum P_0 Q_0}(100) \qquad \textbf{(15.6)}$$

La propuesta de Ernst Louis Étienne Laspeyres (Halle an der Saale, Alemania, 1834-1913) consiste en utilizar las cantidades del año base para calcular las sumas de valores de ambos periodos, por lo que la fórmula se convierte en:

$$IVL = \frac{\sum P_n Q_0}{\sum P_0 Q_0}(100) \qquad \textbf{(15.7)}$$

Nótese que el único cambio es el subíndice de las cantidades del numerador, que pasó de Q_n a Q_0, lo cual indica que se utilizan las mismas cantidades del año base para calcular los valores tanto del periodo n, como los del periodo base, el 0.

Ejemplo 15.4

En el ejemplo 15.3 se calculó, aparte de otros, el índice agregado de valor para el ajonjolí, el arroz palay y el cártamo, para 2008 utilizando 1998 como año base. Calcule ahora el índice de Laspeyres para el mismo año de 2008.

Solución: Se encontró antes que el índice agregado de valor fue:

$$IV = \frac{\sum P_n Q_n}{\sum P_0 Q_0}(100)$$

$$= \frac{9.68(34.32) + 3.63(224.37) + 3.71(95.83)}{4.94(31.65) + 1.64(458.11) + 2.18(171.22)}(100)$$

$$= \frac{1\,502.21}{1\,280.91}(100) = 117.27$$

Calculando ahora el índice de Laspeyres:

$$IVL = \frac{\sum P_n Q_0}{\sum P_0 Q_0}(100)$$

$$= \frac{9.68(31.65) + 3.63(458.11) + 3.71(171.22)}{4.94(31.65) + 1.64(458.11) + 2.18(171.22)}(100)$$

$$= \frac{2\,604.54}{1\,280.91}(100) = 203.34$$

Que es un índice de casi el doble del obtenido antes y que refleja sobre todo las mayores cantidades de producción que se dieron en el año base para el arroz y el cártamo, 458.11 y 171.22 respectivamente, en tanto que en el índice de valor simple calculado antes utilizando las cantidades de 2008, se utilizaron valores mucho menores: 224.37 y 95.83.

15.4.2 Índice de Paasche

Índice de Paasche. Se calcula utilizando, tanto en el numerador como en el denominador, las cantidades del año de interés.

El índice de Paasche (Herman Paasche, Magdeburg, Alemania, 1851, Detroit, EUA, 1925) se distingue porque se calcula utilizando, tanto en el numerador como en el denominador, las cantidades del año de interés, por lo que la fórmula es:

$$IVP = \frac{\sum P_n Q_n}{\sum P_0 P_n}(100) \qquad \textbf{(15.8)}$$

Ejemplo 15.5

Utilice de nuevo los datos de los ejemplos 15.3 y 15.4, calcule ahora el índice de valor de Paasche para los datos del ajonjolí, el arroz y el cártamo.

Solución: Se reproducen de nuevo las operaciones realizadas para calcular el índice de valor del ejemplo 15.3:

$$IV = \frac{\sum P_n Q_n}{\sum P_0 P_0}(100)$$

$$= \frac{9.68(34.32) + 3.63(224.37) + 3.71(95.83)}{4.94(31.65) + 1.64(458.11) + 2.18(171.22)}(100)$$

$$= \frac{1\,502.21}{1\,280.91}(100) = 117.27$$

Como en el índice de valor de Paasche se utilizan las cantidades del periodo de interés, n, 2008, en este caso:

$$IVP = \frac{\sum P_n Q_n}{\sum P_0 Q_n}(100) = \frac{9.68(34.32) + 3.63(224.37) + 3.71(95.83)}{4.94(34.32) + 1.64(224.37) + 2.18(95.83)}(100)$$

$$= \frac{1\,502.21}{746.417}(100) = 224.86$$

que, al igual que el índice de Laspeyres, es otra vez de aproximadamente el doble del obtenido antes.

Este índice de Paasche tiene la ventaja de basarse en pautas actuales de consumo, en tanto que el de Laspeyres se basa en pautas de consumo pasadas.

Índice ideal de Fischer. Es la media geométrica de los índices de Laspeyres y de Paasche.

15.4.3 Índice ideal de Fischer

Este índice es la media geométrica[1] de los índices de Laspeyres y de Paasche, por lo que la fórmula que resume la manera de calcularlo es:

$$IVF = \sqrt{\left(\frac{\sum P_n Q_0}{\sum P_0 Q_0}\right)\left(\frac{\sum P_n Q_n}{\sum P_0 Q_n}\right)}(100) \qquad \textbf{(15.9)}$$

EJEMPLO 15.6

Calcule el índice ideal de Fischer para 2008 con los datos de ajonjolí, arroz y cártamo, utilizando 1998 = 100.

Solución: Utilice los resultados obtenidos antes en los ejemplos 15.4 y 15.5, sustituya en la fórmula (15.9):

$$IVF = \sqrt{\left(\frac{\sum P_n Q_0}{\sum P_0 Q_0}\right)\left(\frac{\sum P_n Q_n}{\sum P_0 Q_n}\right)}(100) = \sqrt{2.0334(2.2486)}\,(100)$$

$$= \sqrt{4.57230324}\,(100) = 213.83$$

Este índice ideal de Fischer satisface varios criterios que este estadístico propuso como medidas para evaluar números índices en su libro *The Making of Index Numbers: A Study of their Varieties, Tests and Reliabilities*, publicado inicialmente en 1922, entre los que destaca el de la prueba de reversión temporal, mediante la cual se demuestra que se obtiene el mismo resultado, sin importar qué punto del tiempo se tome como base de la comparación, el periodo base, o el periodo *n*.

EJERCICIOS 15.3 y 15.4 Números índice agregados

En la tabla 15.3 se muestran los precios por kilogramo y el volumen de producción de 23 productos agrícolas, cíclicos y perennes, para los años 1990 y 1995-2008, con datos del Instituto Nacional de Estadística y Geografía.[2]

Tabla 15.3 Precios y volumen de producción de 23 productos.

									Cíclicos				
Año	Ajonjolí	Arroz palay	Cártamo	Cebada grano	Fresa	Frijol	Maíz grano	Sorgo grano	Soya	Trigo grano	Chile verde	Tomate rojo (jitomate)	Papa
Volumen de producción (miles de toneladas)													
1990	60	394	159	492	107	1 287	14 635	5 978	575	3 931	851	1 885	1 286
1995	21	367	113	487	106	1 271	18 353	4 170	190	3 468	1 187	1 941	1 269
1996	47	394	182	586	84	1 349	18 026	6 809	56	3 375	1 207	2 010	1 282
1997	21	469	163	471	60	965	17 656	5 712	185	3 657	1 833	1 924	1 317
1998	32	458	171	411	88	1 261	18 455	6 475	150	3 235	1 850	2 257	1 281
1999	31	327	263	454	105	1 059	17 706	5 720	133	3 021	1 800	2 418	1 477
2000	41	351	96	713	114	888	17 557	5 842	102	3 493	1 742	2 086	1 627
2001	43	227	111	762	101	1 063	20 134	6 567	122	3 275	1 896	2 150	1 628
2002	20	227	53	737	97	1 549	19 298	5 206	86	3 236	1 783	1 990	1 483

[1] Para repasar la media o promedio geométrico, véase la sección 3.1.4.

[2] Adaptado de las tablas 12.5 y 12.6 del capítulo 12 de: http://www.inegi.org.mx/est/contenidos/espanol/sistemas/sisnav/default.aspx?proy=aeeum&edi=0000&ent=00, 14 de febrero de 2011.

									Cíclicos				
Año	Ajonjolí	Arroz palay	Cártamo	Cebada grano	Fresa	Frijol	Maíz grano	Sorgo grano	Soya	Trigo grano	Chile verde	Tomate rojo (jitomate)	Papa
Volumen de producción (miles de toneladas)													
2003	31	273	201	1082	122	1415	20701	6759	126	2716	1777	2171	1662
2004	33	279	231	932	144	1 163	21 686	7 004	133	2 321	1 865	2 315	1 507
2005	20	291	94	761	129	827	19 339	5 524	187	3 015	2 023	2 246	1 635
2006	21	337	74	869	155	1 386	21 893	5 519	81	3 378	2 077	2 093	1 523
2007	29	295	113	653	140	994	23 513	6 203	88	3 515	2 259	2 425	1 751
2008	34	224	96	781	178	1 111	24 410	6 593	153	4 214	2 052	2 263	1 670
Precio de producción (kg)													
1990	2.02	0.55	0.67	0.56	1.45	1.99	0.61	0.34	0.82	0.51	1.26	0.78	0.59
1995	5.11	1.07	1.32	1.00	1.82	2.19	1.09	0.94	1.47	0.90	2.12	1.32	1.77
1996	4.60	1.62	1.99	1.42	2.69	4.26	1.43	1.14	2.13	1.77	2.44	2.26	2.43
1997	4.33	1.52	2.08	1.38	3.43	5.47	1.35	0.98	2.23	1.32	3.46	3.44	2.16
1998	4.94	1.64	2.18	1.44	4.30	6.04	1.45	1.02	2.29	1.37	3.87	4.12	3.18
1999	5.73	1.78	1.95	1.44	4.81	5.25	1.45	0.98	2.46	1.37	3.62	3.73	3.64
2000	5.66	1.47	1.61	1.48	5.47	5.22	1.51	1.05	1.80	1.47	4.21	3.84	3.32
2001	4.96	1.48	1.31	1.63	6.37	6.25	1.45	0.99	1.85	1.22	3.64	3.05	3.29
2002	5.17	1.64	1.78	1.57	7.64	5.73	1.50	1.19	2.03	1.21	3.43	3.12	4.40
2003	6.25	1.66	2.27	1.65	7.21	5.08	1.62	1.30	3.00	1.42	4.16	4.23	4.38
2004	7.68	1.82	2.35	1.79	6.72	5.73	1.68	1.33	2.72	1.66	5.91	6.21	4.34
2005	7.95	1.90	2.26	1.78	9.47	6.90	1.58	1.20	2.30	1.62	4.87	4.41	4.62
2006	6.99	1.91	2.34	1.91	8.96	6.30	2.01	1.57	2.61	1.68	3.87	5.88	4.78
2007	8.00	2.08	2.36	2.17	7.76	6.98	2.44	1.92	3.64	2.07	5.31	4.75	4.43
2008	9.68	3.63	3.71	3.26	6.89	9.16	2.82	2.31	4.55	3.68	5.50	5.61	4.70

	Perennes									
	Agua-cate	Café cereza	Durazno	Fresa	Mango	Manzana	Naranja	Limón	Plátano	Uva
Volumen de la producción (miles de toneladas)										
1990	686	1 641	161	ND	1 074	457	2 220	685	1 986	429
1995	790	1 726	120	25	1 342	413	3 572	947	2 033	476
1996	838	1 976	151	36	1 190	427	3 985	1 089	2 210	408
1997	762	1 852	129	39	1 501	629	3 944	1 096	1 714	473
1998	877	1 507	116	31	1 474	370	3 331	1 171	1 526	478
1999	879	1 641	126	32	1 508	450	3 520	1 347	1 752	483
2000	907	1 837	147	27	1 559	338	3 813	1 640	1 871	372
2001	940	1 646	176	30	1 577	443	4 035	1 573	2 114	436
2002	901	1 700	198	46	1 523	480	4 020	1 706	1 997	363
2003	905	1 622	169	24	1 362	495	3 846	1 749	2 066	331

(*continúa*)

Tabla 15.3 (*continuación*)

	Agua-cate	Café cereza	Durazno	Fresa	Mango	Manzana	Naranja	Limón	Plátano	Uva
Volumen de la producción (miles de toneladas)										
2004	987	1 697	202	33	1 573	573	3 977	1 913	2 361	305
2005	1 022	1 599	208	34	1 368	584	4 113	1 792	2 250	332
2006	1 134	1 519	222	37	1 735	602	4 157	1 852	2 196	244
2007	1 143	1 459	192	36	1 643	505	4 249	1 923	1 965	356
2008	1 162	1 415	191	31	1 717	512	4 297	2 229	2 151	266
Precio de la producción (kg)										
1990	1.49	0.89	1.40	ND	0.84	0.84	0.42	0.50	0.46	1.30
1995	1.39	2.23	3.05	1.80	1.30	1.85	0.56	1.04	0.98	2.20
1996	2.11	3.08	3.78	4.31	1.51	2.58	0.70	1.15	1.18	2.77
1997	4.27	3.84	4.41	3.69	1.41	1.59	0.59	1.38	1.24	3.86
1998	3.90	4.20	4.60	5.36	1.89	3.37	0.78	1.64	1.68	4.98
1999	8.38	4.19	5.77	5.69	2.13	3.54	1.09	2.13	1.78	4.27
2000	4.65	2.88	5.77	5.17	1.93	3.52	0.79	2.01	1.61	5.08
2001	5.35	1.86	5.07	5.00	1.96	2.81	0.61	1.40	1.81	5.79
2002	4.48	1.62	4.33	6.50	2.35	3.09	0.71	1.41	1.32	8.51
2003	5.94	1.83	5.48	6.46	2.31	3.34	0.89	1.74	1.20	10.82
2004	6.16	1.69	5.53	7.04	2.17	3.55	0.78	1.63	1.44	7.86
2005	7.46	2.26	5.75	6.78	2.49	3.59	0.65	1.44	1.77	9.13
2006	8.04	2.67	5.93	9.69	2.29	4.72	0.93	1.78	1.73	10.85
2007	10.52	3.34	6.17	10.00	2.50	5.62	1.07	2.13	2.66	12.68
2008	10.72	3.92	6.20	8.28	2.20	5.32	0.95	2.17	2.10	13.00

Fuente: Adaptado de http://www.inegi.org.mx/est/contenidos/espanol/sistemas/sisnav/default.aspx?proy=aeeum&edi=0000&ent=00, 1 febrero 2011, 22 de febrero de 2012.

1. Calcule los índices simples de precio, cantidad y valor para 2008 y para cada uno de los 23 artículos. Considere lo siguiente:

a) 1990 = 100.
b) 1995 = 100.
c) Interprételos.

2. Calcule los índices agregados de precio, cantidad y valor para 2008 y para:

a) Productos cíclicos.
  i) 1990 = 10.
  ii) 1995 = 10.0.

b) Productos perennes
  i) 1990 = 100.
  • 1995 = 100.

c) Explique qué significan.

3. Una ensambladora de radios portátiles compra una misma pieza a 3 proveedores diferentes, los cuales tienen precios unitarios y suministran cantidades diferentes. En la siguiente tabla se muestran los datos referentes al 2008 y 2011. Calcule:

a) Índice de precios.
b) Índice de valor, utilizando en ambos casos 2008 = 100.

Proveedor	Cantidad	Precio 2008	Precio 2011
Metales Ramírez	200	6.50	7.05
Grupo Al	250	6.65	7.00
MetalMex	170	6.55	7.25

4. En la siguiente tabla se muestran los precios y cantidades de los 2 principales artículos de una empresa productora de artículos de limpieza para el hogar, en 2007 y 2010.

a) Calcule el índice agregado de valor.
b) Determine el índice de Laspeyres.
c) Calcule el índice de Paasche.
d) Calcule el índice ideal de Fischer, tomando 2007 = 100.

Artículo	Precio		Cantidad	
	2007	2010	2007	2010
Limpiador de pisos	47	57.5	1700	2100
Lavatrastes	17	28	2450	3300

5. Se registraron los precios unitarios de algunos alimentos que fueron adquiridos en una muestra pequeña de familias durante los meses de agosto y septiembre de 2008, se estableció una ponderación de 10 en 10 unidades con base en lo que las familias consideraron que consumieron más (10 para el menor consumo y 40 para el mayor consumo). Los resultados se muestran a continuación.

	Precio		Cantidad
	Ago.	Sept.	
Carne de res (kg)	67.50	79.90	30
Carne de pollo (kg)	18.00	26.80	40
Jamón (kg)	54.70	60.00	20
Salchicha (kg)	37.40	32.60	10

Fuente: http://www.sedeco.df.gob.mx/indicadores/abasto/cbasica/index2008.html/

a) Calcule índice de precios agregado.
b) Obtenga el índice de precios ponderado, tomando agosto como base para ambos casos.

6. Una empresa refresquera vende 3 bebidas de diferentes sabores. Los precios unitarios de 2004 y 2007 y cajas vendidas en 2007 se muestran a continuación.

Sabor	Precio		Cantidad
	2004	2007	2007
Toronja	3.50	5.50	7 000
Naranja	4.50	5.00	35 000
Manzana	5.00	6.00	60 000

a) Calcule el índice de precios no ponderado.
b) Obtenga el índice de precios ponderado, para 2007 ambos.

7. Una fábrica produce 3 tipos de piezas de refacción para autos. En la siguiente tabla se muestran los precios unitarios a las que se vendieron durante los meses de mayo y agosto de 2008, así como las cantidades en que surtieron a los clientes.

a) Calcule el índice de precios ponderado.
b) Determine el índice de Laspeyres.
c) Calcule el índice de Paasche para agosto.

	Precio		Cantidad	
Refacción	Mayo	Ago.	Mayo	Ago.
A	223	239	438	502
B	354	430	153	176
C	543	600	200	100

15.5 Números índices en cadena

Los índices en cadena se obtienen calculando índices de periodos consecutivos en una serie de datos.

Ejemplo 15.7

En la tabla 15.4 se muestran datos de precios mensuales de cebada de enero de 2009 a junio de 2011, junto con los índices en cadena.

Tabla 15.4 Precios internacionales de la cebada e índices en cadena.

Periodo	Cebada (USD por tonelada métrica)	Índices en cadena
Ene. 2009	121.61	
Feb. 2009	112.57	92.57
Mar. 2009	114.94	102.11
Abr. 2009	111.02	96.59
Mayo 2009	129.39	116.55
Jun. 2009	148.68	114.91
Jul. 2009	139.96	94.14
Ago. 2009	122.3	87.38
Sep. 2009	103.53	84.65
Oct. 2009	130.55	126.10
Nov. 2009	155.26	118.93
Dic. 2009	150.77	97.11
Ene. 2010	146.6	97.23
Feb. 2010	137.3	93.66
Mar. 2010	147.06	107.11

(*continúa*)

Tabla 15.4 (*continuación*)

Periodo	Cebada (USD por tonelada métrica)	Índices en cadena
Abr. 2010	151.71	103.16
Mayo 2010	142.95	94.23
Jun. 2010	145.97	102.11
Jul. 2010	156.36	107.12
Ago. 2010	161.03	102.99
Sep. 2010	168.2	104.45
Oct. 2010	174.61	103.81
Nov. 2010	179.12	102.58
Dic. 2010	189.6	105.85
Ene. 2011	195.12	102.91
Feb. 2011	196.37	100.64
Mar. 2011	202.53	103.14
Abr. 2011	208.7	103.05
Mayo 2011	208.72	100.01
Jun. 2011	210.17	100.69

Fuente: http://www.indexmundi.com/commodities/, consultado el 20 de febrero de 2012.

El cálculo de los índices en cadena es sumamente sencillo: el primero que aparece en la tabla 15.4 se encontró dividendo el precio de febrero de 2009 entre el de enero de ese mismo año y multiplicando ese cociente por 100; los demás se calcularon de la misma manera y es fácil ver la forma en la que se pueden calcular esos índices en cadena con Excel. Una utilidad de estos índices en cadena salta a la vista cuando se les grafica junto con los precios originales, como en la figura 15.1.

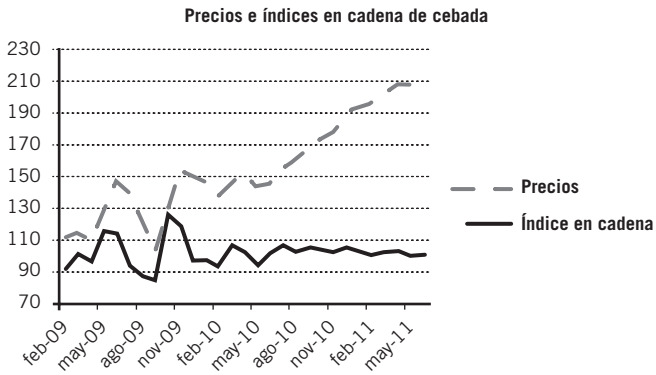


Figura 15.1 Gráficas de los precios y los índices en cadena de cebada.

La gráfica de precios muestra cómo crecieron los precios hasta llegar a los 210 dólares en junio de 2011, en tanto que la gráfica de los índices en cadena muestra que los precios subieron y bajaron drásticamente entre febrero y diciembre de 2009 para, después, subir en forma constante hasta el fin del periodo.

15.5.1 Números índice en cadena y rendimientos bursátiles

En el análisis de instrumentos que cotizan en las bolsas de valores es común que se calculen los números índice en cadena para analizar, con diversos propósitos, el comportamiento de sus rendimientos.

EJEMPLO 15.8

En la tabla 15.5 se muestran los precios de las acciones de Cemex CPO, junto con los números índices en cadena calculados, al igual que antes, dividiendo cada precio entre el precio anterior y multiplicando el cociente por 100. Dichos índices en cadena de los precios de las acciones contienen ya los rendimientos, pues para calcularlos basta con restarle 100 a cada índice, como se ve en la cuarta columna de la tabla 15.5. Estos rendimientos diarios que son, en última instancia, números índices en cadena, se utilizan en diversos análisis financieros bursátiles.

Tabla 15.5 Precios de Cemex CPO, índices en cadena y rendimientos

Fecha	Precio	Índice en cadena	Rendimiento diario
22 jul. 2011	8.75		
25 jul. 2011	8.6	98.29	–1.71
26 jul. 2011	8.59	99.88	–0.12
27 jul. 2011	8.42	98.02	–1.98
28 jul. 2011	7.95	94.42	–5.58
29 jul. 2011	8.3	104.40	4.40
1 ago. 2011	8.02	96.63	–3.37
2 ago. 2011	7.49	93.39	–6.61
3 ago. 2011	7.33	97.86	–2.14
4 ago. 2011	6.84	93.32	–6.68
5 ago. 2011	6.93	101.32	1.32
8 ago. 2011	6.17	89.03	–10.97
9 ago. 2011	6.49	105.19	5.19
10 ago. 2011	6.2	95.53	–4.47
11 ago. 2011	6.47	104.35	4.35
12 ago. 2011	6.37	98.45	–1.55
15 ago. 2011	6.83	107.22	7.22
16 ago. 2011	6.96	101.90	1.90
17 ago. 2011	6.79	97.56	–2.44
18 ago. 2011	6.37	93.81	–6.19
19 ago. 2011	6.15	96.55	–3.45

Fecha	Precio	Índice en cadena	Rendimiento diario
22 ago. 2011	6.1	99.19	–0.81
23 ago. 2011	6.37	104.43	4.43
24 ago. 2011	6.5	102.04	2.04
25 ago. 2011	6.38	98.15	–1.85

Fecha	Precio	Índice en cadena	Rendimiento diario
26 ago. 2011	6.26	98.12	–1.88
29 ago. 2011	6.68	106.71	6.71
30 ago. 2011	6.6	98.80	–1.20
31 ago. 2011	6.65	100.76	0.76

ejercicios 15.5 Números índices en cadena

En la tabla siguiente se muestran los volúmenes de producción de diversos productos agrícolas.

Año	Cíclicos												
	Ajonjolí	Arroz palay	Cártamo	Cebada grano	Fresa	Frijol	Maíz grano	Sorgo grano	Soya	Trigo grano	Chile verde	Tomate rojo (jitomate)	Papa
	Volumen de la producción (miles de toneladas)												
1990	60	394	159	492	107	1 287	14 635	5 978	575	3 931	851	1 885	1 286
1995	21	367	113	487	106	1 271	18 353	4 170	190	3 468	1 187	1 941	1 269
1996	47	394	182	586	84	1 349	18 026	6 809	56	3 375	1 207	2 010	1 282
1997	21	469	163	471	60	965	17 656	5 712	185	3 657	1 833	1 924	1 317
1998	32	458	171	411	88	1 261	18 455	6 475	150	3 235	1 850	2 257	1 281
1999	31	327	263	454	105	1 059	17 706	5 720	133	3 021	1 800	2 418	1 477
2000	41	351	96	713	114	888	17 557	5 842	102	3 493	1 742	2 086	1 627
2001	43	227	111	762	101	1 063	20 134	6 567	122	3 275	1 896	2 150	1 628
2002	20	227	53	737	97	1 549	19 298	5 206	86	3 236	1 783	1 990	1 483
2003	31	273	201	1 082	122	1 415	20 701	6 759	126	2 716	1 777	2 171	1 662
2004	33	279	231	932	144	1 163	21 686	7 004	133	2 321	1 865	2 315	1 507
2005	20	291	94	761	129	827	19 339	5 524	187	3 015	2 023	2 246	1 635
2006	21	337	74	869	155	1 386	21 893	5 519	81	3 378	2 077	2 093	1 523
2007	29	295	113	653	140	994	23 513	6 203	88	3 515	2 259	2 425	1 751
2008	34	224	96	781	178	1 111	24 410	6 593	153	4 214	2 052	2 263	1 670

Fuente:www.inegi.org.mx/est/contenidos/espanol/sistemas/sisnav/default.aspx?proy=aeeum&edi=0000&ent=00, 1 febrero 2011.

1. Calcule los índices en cadena del ajonjolí, grafíquelos y comente.
2. Calcule los índices en cadena de la fresa, grafíquelos y comente.
3. Calcule los índices en cadena del maíz en grano, grafíquelos y comente.
4. Calcule los índices en cadena de la soya, grafíquelos y comente.
5. Calcule los índices en cadena del chile verde, grafíquelos y comente.
6. Compare los 5 índices en cadena anteriores y comente.

En la tabla siguiente se muestran los precios de cereales en el mercado mundial, de junio de 2006 a junio de 2011.

	Cebada	Maíz	Arroz	Sorgo	Trigo
Mes	USD por tonelada métrica				
Jun. 2006	106.88	109.55	305.59	111.88	195.16
Jul. 2006	112.1	114.24	312.43	119.99	202.43
Ago. 2006	114.53	115.21	313.39	114.44	189.91
Sep. 2006	115.44	120.26	312.9	119.55	195.98
Oct. 2006	136.6	142.17	309.64	138.23	212.09
Nov. 2006	148.15	164.08	300.59	167.09	205.81
Dic. 2006	150.97	160.66	309.29	170.61	204.31
Ene. 2007	152.13	165.1	313.48	175.05	196.07

(*continúa*)

(*continuación*)

	Cebada	Maíz	Arroz	Sorgo	Trigo
Mes	USD por tonelada métrica				
Feb. 2007	151.03	177.35	316.05	180.64	199.98
Mar. 2007	157.06	169.52	326.18	169.96	199.1
Abr. 2007	157.12	152.58	322.29	149.52	198.31
Mayo 2007	161.69	156.44	320.61	150.01	195.72
Jun. 2007	184.45	164.5	326.29	154.79	223.04
Jul. 2007	177.91	147.13	332.55	138.47	238.41
Ago. 2007	158.88	151.01	331.48	150.28	259.73
Sep. 2007	184.66	160.05	330	163.31	326.54
Oct. 2007	197.11	164.09	335.3	163.19	335.15
Nov. 2007	187.61	171.06	356.5	170.1	321.81
Dic. 2007	198.93	180.25	378	187.01	368.62
Ene. 2008	205.61	206.53	393.48	212.67	369.59
Feb. 2008	216.54	219.95	481.14	218.5	425
Mar. 2008	228.66	233.85	672.64	224.93	439.72
Abr. 2008	237.24	246.67	1 015.21	240.28	362.23
Mayo 2008	238.31	243.46	1 009.32	238.24	328.76
Jun. 2008	240.98	287.11	834.6	262.19	348.55
Jul. 2008	248.31	266.94	799	218.82	328.18
Ago. 2008	212.32	235.16	737	209.34	329.34
Sep. 2008	189.43	233.91	722	216.01	295.55
Oct. 2008	144	182.96	624	163.63	237.38
Nov. 2008	130.16	164.27	563.25	150.75	226.85
Dic. 2008	114.16	158.16	550.75	138.6	220.14
Ene. 2009	121.61	173.24	615.25	153.26	239.36
Feb. 2009	112.57	163.13	634	144.13	224.69
Mar. 2009	114.94	164.52	625.25	138.59	230.98
Abr. 2009	111.02	168.72	577.25	154.14	233.47
Mayo 2009	129.39	180.31	540.75	160.08	256.7
Jun. 2009	148.68	178.83	548.6	153.04	253.41
Jul. 2009	139.96	151.76	623	133.8	224.95
Ago. 2009	122.3	152.01	576.25	142.37	210.37
Sep. 2009	103.53	150.57	593.67	141.83	191.09
Oct. 2009	130.55	167.22	566.25	159.05	198.85
Nov. 2009	155.26	171.61	566.25	166.01	211.04
Dic. 2009	150.77	164.58	606	166.35	206.25
Ene. 2010	146.6	167.21	598	161.79	201.51
Feb. 2010	137.3	161.63	584.75	154.06	194.54
Mar. 2010	147.06	159.01	540.13	154.7	191.07
Abr. 2010	151.71	157.66	502.23	149.41	192.82
Mayo 2010	142.95	163.77	472.48	147.35	181.88
Jun. 2010	145.97	152.87	458.55	130.98	157.67
Jul. 2010	156.36	163.92	470.68	132.4	195.82
Ago. 2010	161.03	175.6	486.86	143.4	246.25
Sep. 2010	168.2	205.84	519.95	184.9	271.69
Oct. 2010	174.61	235.7	533.13	201.04	270.23
Nov. 2010	179.12	238.24	543.14	203.25	274.08
Dic. 2010	189.6	250.63	536.78	221.57	306.53
Ene. 2011	195.12	265.29	528.38	246.32	326.55
Feb. 2011	196.37	293.4	532.8	253.15	348.15
Mar. 2011	202.53	290.43	508.96	266.14	316.75
Abr. 2011	208.7	318.74	500.57	289.61	336.12
Mayo 2011	208.72	308.58	500.55	261.32	354.47
Jun. 2011	210.17	310.54	515.44	260.43	326.45

http://www.indexmundi.com/commodities/. Revisado el 20 de febrero de 2011.

7. Calcule los índices en cadena de la cebada, grafíquelos y comente.

8. Calcule los índices en cadena del maíz, grafíquelos y comente.

9. Calcule los índices en cadena del arroz, grafíquelos y comente.

10. Calcule los índices en cadena del sorgo, grafíquelos y comente.

11. Calcule los índices en cadena del trigo, grafíquelos y comente.

12. Visite la sección de finanzas del sitio de Internet de Yahoo! (http://mx.finance.yahoo.com/) y obtenga los precios históricos de las acciones de:

a) Alfaa.mx
b) Gfinburo.mx
c) Kimbera.mx
d) Grumab.mx

13. Los nombres anotados en los incisos de la pregunta anterior son las claves con las que Yahoo! identifica las acciones; para cada una de ellas, construya:

a) Índices en cadena de los precios.

b) Índices en cadena de rendimientos.

c) Grafique los índices de los incisos *a*) y *b*).

d) Comente los índices de los incisos *a*) y *b*).

15.6 Índices para propósitos especiales

Hasta ahora se revisaron los procedimientos para calcular los índices de precios, de cantidades y de valor para artículos únicos o para conjuntos de ellos. En esta sección se revisan algunos índices que se consideran medidas importantes de las condiciones económicas y de negocios y son, básicamente, de 2 tipos: índices de precios al consumidor y al productor que se utilizan entre otras cosas para medir la inflación, e índices bursátiles que se emplean para medir el comportamiento de los mercados accionarios bursátiles.

15.6.1 Índices de precios al consumidor y al productor

Los índices de precios al consumidor se utilizan primordialmente para medir los aumentos de precios de un conjunto de artículos representativos del consumo básico de las familias y, como suelen ser representativos, se utilizan ampliamente para medir la inflación, en tanto que el *Índice Nacional de Precios al Productor* (INPP) es un conjunto de indicadores de precios. Su finalidad es proporcionar mediciones sobre la variación de los precios de una canasta fija de bienes y servicios representativa de la producción nacional. El **precio productor** se define como el precio fijado por el productor a la primera instancia compradora de su producto, excluyendo impuestos y costos de transportación facturados por separado.[3]

Precio productor. Precio fijado por el productor a la primera instancia compradora de su producto, excluyendo impuestos y costos de transportación facturados por separado.

15.6.1.1 El Índice Nacional de Precios al Consumidor

En el caso de México, de acuerdo con la definición del Banco de México (Banxico):

> El Índice Nacional de Precios al Consumidor (INPC)[4] es un indicador económico diseñado específicamente para medir el cambio promedio de los precios en el tiempo, mediante una canasta ponderada de bienes y servicios representativa del consumo de las familias urbanas de México.

A dicha canasta se le conoce como *canasta básica*. Como las variaciones del INPC se toman como una aproximación de las variaciones de los precios de los bienes y servicios comerciados en el país, es el indicador oficial de la inflación en México.

Banxico calcula el INPC utilizando la fórmula de Laspeyres (ver la sección 15.4.1), con precios representativos de todas las localidades del país con más de 20 000 habitantes (46 ciudades), incluyendo los diferentes puntos de venta y las distintas marcas, presentaciones y modalidades de los bienes y servicios que se incluyen en la canasta; las categorías en las que se divide la canasta básica, según el objeto del gasto son:

- Alimentos, bebidas y tabaco.
- Ropa, calzado y accesorios.
- Vivienda.
- Muebles, aparatos y accesorios domésticos.
- Salud y cuidado personal.
- Transporte.
- Educación y esparcimiento.
- Otros servicios.

[3] http://www.banxico.org.mx/politica-monetaria-e-inflacion/material-de-referencia/intermedio/inflacion/elaboracion-del-inpp.html, 22 de febrero de 2011.

[4] Esta sección fue elaborada en parte con información del Banco de México, al 10 de febrero de 2011: http://www.banxico.org.mx/polmoneinflacion/didactico/preguntasFrecuentes/PreguntasFrecuentesINPC.html

Banxico publica un documento detallado sobre el procedimiento para calcular el INPC[5] que incluye, entre otras secciones, la definición de la canasta y el método de cálculo.

El INPC se calcula quincenal y mensualmente y se ofrecen series de datos quincenales y mensuales. Las series históricas de este INPC se pueden consultar en http://www.inegi.org.mx/est/contenidos/proyectos/inp/inpc.aspx

Aquí se debe comentar que, aunque parte de la información de esta sección se elaboró en febrero de 2011, de acuerdo con la Ley del Sistema Nacional de Información Estadística y Geográfica, publicada en el *Diario Oficial de la Federación* el 16 de abril de 2008, a partir del 15 de julio de 2011 se transfirió del Banco de México al Instituto Nacional de Geografía y Estadística (INEGI) la facultad exclusiva de elaborar y publicar los índices de precios. Como esa fecha de julio de 2011 transcurrió mientras se elaboraba este material, los datos sobre el INPC se obtuvieron ya del sitio del INEGI.

En la tabla 15.6 se muestran los valores del INPC de enero de 2008 a septiembre de 2011, obtenidos del sitio del INEGI mencionado y que tienen como base diciembre de 2010.

Tabla 15.6 INPC mensual enero 2008-septiembre 2011.

Ene. 2008	87.01013	Dic. 2009	95.64384
Feb. 2008	87.3763	Ene. 2010	96.79132
Mar. 2008	87.9666	Feb. 2010	97.30024
Abr. 2008	88.14244	Mar. 2010	97.90226
Mayo 2008	88.06107	Abr. 2010	97.42299
Jun. 2008	88.38725	Mayo 2010	96.89545
Jul. 2008	88.95823	Jun. 2010	96.88097
Ago. 2008	89.4775	Jul. 2010	97.13129
Sep. 2008	90.05813	Ago. 2010	97.4092
Oct. 2008	90.65601	Sep. 2010	97.91329
Nov. 2008	91.70074	Oct. 2010	98.53737
Dic. 2008	92.36413	Nov. 2010	99.29248
Ene. 2009	92.40689	Dic. 2010	100
Feb. 2009	92.75237	Ene. 2011	100.282
Mar. 2009	93.34405	Feb. 2011	100.72
Abr. 2009	93.55851	Mar. 2011	100.824
Mayo 2009	93.25371	Abr. 2011	100.846
Jun. 2009	93.45507	Mayo 2011	100.001
Jul. 2009	93.69988	Jun. 2011	100.135
Ago. 2009	93.98813	Jul. 2011	100.592
Sep. 2009	94.37499	Ago. 2011	100.682
Oct. 2009	94.60255	Sep. 2011	100.958
Nov. 2009	95.18733		

Fuente: http://www.inegi.org.mx/sistemas/indiceprecios/Estructura.aspx?idEstructura=R6500100010&T=Índices de Precios al Consumidor&ST=Principales índices, consultado el 19 de febrero de 2012.

[5] http://www.banxico.org.mx/politica-monetaria-e-inflacion/material-de-referencia/intermedio/inflacion/elaboracion-inpc/%7B50ECE064-0F0A-F533-1477-3C77A959CE7B%7D.pdf (19 de febrero de 2011").

Dos de las principales aplicaciones del INPC son medir la inflación y deflacionar series de tiempo, es decir, reexpresar series de tiempo descontándoles la inflación. En las 2 subsecciones siguientes se ilustran estos procedimientos.

15.6.1.2 Cálculo de la inflación mediante el INPC

Esta aplicación es sumamente sencilla; como los índices de precios al consumidor son, precisamente, índices de precios, la inflación porcentual se calcula simplemente dividiendo los índices de los periodos de interés.

EJEMPLO 15.9

Utilizando los datos del INPC que se resumen en la tabla 15.4 se puede calcular la inflación de enero de 2010 a enero de 2011 de la manera que se ilustra a continuación.

Solución: Los INPC de esos 2 meses de enero fueron: 100.282 para enero de 2011, y 96.79132 para enero de 2010:

$$Inflación = \frac{100.282}{96.79132}(100) = 103.606$$

Resultado que indica que los precios de la canasta básica con la que se calcula el INPC subieron aproximadamente 3.61% en un año de enero de 2010 a enero de 2011.

15.6.1.3 Cambio de periodo base

El cambio de periodo que se usa como base para calcular números índice puede servir a dos propósitos principales. En primer lugar, puede ser necesario actualizar la base de una serie de números índices que se haya hecho muy vieja. Se suele hacer esto en las series de índices de precios que se construyen constantemente para propósitos estadísticos.

Un segundo propósito consiste en cambiar el periodo base de una serie de números índices para que coincida con el periodo base de otra con la que se desean hacer comparaciones u operaciones y es el caso del procedimiento de deflación de series que se ilustra en la sección siguiente y que, como se verá, es de enorme utilidad. Para cambiar la base de un conjunto de números índice basta con dividir todos ellos entre el valor del índice para el periodo que se desea usar como base.

EJEMPLO 15.10

En la tabla 15.6 se listan los INPC mensuales de enero de 2008 a septiembre de 2011, con base en diciembre de 2010, que es cuando el INPC tiene precisamente el valor de 100. Cambie la base de esta serie de números índice a enero de 2008.

Solución: Tal como se anota antes, para cambiar la base a enero 2008 = 100, basta con dividir todos ellos entre 87.01013, que es el valor del índice para el año base que se desea. En la tabla 15.7 se resumen los resultados.

Tabla 15.7 INPC mensual enero 2008-enero 2011.

Periodo	INPC Dic 2011 = 100	INPC Enero 2008 = 100
Ene. 2008	87.01013	100.00000
Feb. 2008	87.3763	100.42084
Mar. 2008	87.9666	101.09926
Abr. 2008	88.14244	101.30136
Mayo 2008	88.06107	101.20784
Jun. 2008	88.38725	101.58271
Jul. 2008	88.95823	102.23894
Ago. 2008	89.4775	102.83572
Sep. 2008	90.05813	103.50304
Oct. 2008	90.65601	104.19018
Nov. 2008	91.70074	105.39088
Dic. 2008	92.36413	106.15331
Ene. 2009	92.40689	106.20245
Feb. 2009	92.75237	106.59951
Mar. 2009	93.34405	107.27951
Abr. 2009	93.55851	107.52600
Mayo 2009	93.25371	107.17569
Jun. 2009	93.45507	107.40711
Jul. 2009	93.69988	107.68847

(*continúa*)

Tabla 15.7 (*continuación*)

Periodo	INPC Dic 2011 = 100	INPC Enero 2008 = 100
Ago. 2009	93.98813	108.01975
Sep. 2009	94.37499	108.46437
Oct. 2009	94.60255	108.72591
Nov. 2009	95.18733	109.39798
Dic. 2009	95.64384	109.92265
Ene. 2010	96.79132	111.24144
Feb. 2010	97.30024	111.82634
Mar. 2010	97.90226	112.51823
Abr. 2010	97.42299	111.96741
Mayo 2010	96.89545	111.36111
Jun. 2010	96.88097	111.34447
Jul. 2010	97.13129	111.63216
Ago. 2010	97.4092	111.95156
Sep. 2010	97.91329	112.53091
Oct. 2010	98.53737	113.24816
Nov. 2010	99.29248	114.11600
Dic. 2010	100	114.92915
Ene. 2011	100.282	115.25325
Feb. 2011	100.72	115.75664
Mar. 2011	100.824	115.87616
Abr. 2011	100.846	115.90145
Mayo 2011	100.001	114.93030
Jun. 2011	100.135	115.08430
Jul. 2011	100.592	115.60953
Ago. 2011	100.682	115.71296
Sep. 2011	100.958	116.03017

Este proceso de cambiar el periodo base de un conjunto de números índice no altera la relación que existe entre ellos, pero sí sirve para los 2 propósitos enunciados antes: actualizar bases muy antiguas e igualar las bases de diferentes series para efectos de comparación o combinación, como en la deflación de series que se ilustra en el apartado siguiente.

Ejemplo 15.11

En el ejemplo 15.9 anterior se calculó la inflación de enero de 2010 a enero de 2011 con los datos del INPC de la tabla 15.6. Para mostrar que el cambio de base no altera la relación entre los índices, a continuación se calculará la inflación para ese mismo periodo pero ahora utilizando los números índices con enero de 2008 como periodo base.

Solución: Los INPC durante esos 2 meses de enero fueron 1.152185797726 para enero de 2011 y 111.24144, para enero de 2010, ya que

$$\textit{Inflación} = \frac{115.25325}{111.24144}(100) = 103.606$$

que es el mismo resultado que se encontró antes.

15.6.1.4 Deflación de series de tiempo con el INPC

Otra aplicación importante del INPC es la deflación de series de tiempo, que consiste en descontar la inflación de alguna serie para obtener *valores reales* en vez de los *valores corrientes* que se tienen si no se descuenta la inflación. Estos términos de valores reales y valores corrientes son de uso común en áreas como economía y finanzas, de manera que vale la pena destacar su significado. Los **valores corrientes** o pesos corrientes son los valores al momento en que se utilizan (sin descontar inflación), en tanto que los **valores reales** o pesos reales son esos valores pero después de que se les descontó la inflación.

Valores corrientes. Se utilizan al momento sin descontar la inflación.
Valores reales. Son los valores corrientes ya con el ajuste de la inflación.

La gran mayoría de las series de datos empresariales y económicos, como las ventas, inventarios, etc., se miden en moneda. Estas series muestran con frecuencia un crecimiento al paso del tiempo, lo que suele interpretarse como un incremento en el volumen físico relacionado con la actividad a la que se atribuye la serie. En consecuencia, en periodos en donde los cambios de precios son grandes, los cambios monetarios pueden no reflejar fielmente los cambios en la cantidad, por lo que se utiliza este procedimiento de deflación para ajustar la serie de tiempo de manera que elimine el efecto del cambio en el precio, es decir, de la inflación.

Ejemplo 15.12

En la tabla siguiente, la 15.8, se muestran los salarios mínimos por cada una de las 3 zonas económicas en las que se divide México, de 2000 a 2011, junto con los INPC para los meses de enero de esos mismos años. Deflacione los salarios de la zona A.

Tabla 15.8 Salarios mínimos históricos en México.

Salarios mínimos				INPC (enero)
Vigencia	Zona A	Zona B	Zona C	(Dic 2010 = 100)
01 ene. 2000	37.9	35.1	32.7	59.96995
01 ene. 2001	40.35	37.95	35.85	64.68080
01 ene. 2002	42.15	40.1	38.3	67.88722
01 ene. 2003	43.65	41.85	40.3	71.26327
01 ene. 2004	45.24	43.73	42.11	74.36092
01 ene. 2005	46.8	45.35	44.05	77.60959
01 ene. 2006	48.67	47.16	45.81	80.74517
01 ene. 2007	50.57	49.00	47.60	83.96420
01 ene. 2008	52.59	50.96	49.5	87.01013
01 ene. 2009	54.8	53.26	51.95	92.40689
01 ene. 2010	57.46	55.84	54.47	96.79132
01 ene. 2011	59.82	58.13	56.7	100.282

Fuente: Con datos de www.inegi.org.mx y www.sat.gob.mx

Solución: Se puede ver en la tabla anterior que los salarios mínimos subieron constantemente en esos años. Si se consideran solamente los índices de crecimiento de esos salarios nominales para la zona A se obtiene:

Vigencia	Zona A	Índice
01 ene. 2000	37.9	
01 ene. 2001	40.35	106.4644
01 ene. 2002	42.15	104.461
01 ene. 2003	43.65	103.5587
01 ene. 2004	45.24	103.6426
01 ene. 2005	46.8	103.4483
01 ene. 2006	48.67	103.9957
01 ene. 2007	50.57	103.9038
01 ene. 2008	52.59	103.9945
01 ene. 2009	54.8	104.2023
01 ene. 2010	57.46	104.854
01 ene. 2011	59.82	104.1072

Por ejemplo, el valor 106.4644 se obtuvo dividiendo el salario para ese año de 2001 (40.35), entre el de 2000 (37.9) y multiplicando el cociente por 100, tal como se hace para calcular números índice simples.

Estos índices muestran que los salarios mínimos de la zona A se incrementaron constantemente, cada año, entre 3.44 y 6.46%. Sin embargo, estos cálculos no reflejan el efecto de la inflación sobre los salarios. Es decir, no toman en cuenta el poder de compra que significa ese salario respecto a los precios del mercado. Es por eso que, para tener un panorama claro en cuanto a los aumentos en el salario mínimo, es necesario compararlo con el INPC que es el indicador que nos informa acerca del poder adquisitivo del dinero; en otras palabras, es necesario deflactar la serie de tiempo para obtener los salarios mínimos reales para cada año de la serie.

El proceso de deflación consiste en dividir cada uno de los salarios mínimos entre el Índice Nacional de Precios al Consumidor para ese mismo año pero, en este caso, tenemos que la base de los INPC es diciembre de 2010, por lo que, antes de deflacionar, es necesario cambiar la base de estos INPC a 2000 para que coincida con el primer dato de los salarios mínimos utilizando el procedimiento que se explicó en la sección anterior y que consiste, básicamente, en dividir todos los valores del INPC entre el valor de 2000, ya que se trata de hacer 2000 = 100 y dividir 59.96995 entre sí mismo y multiplicar el cociente por 100, con lo que se obtiene que el INPC para el año 2000 es igual a 100; para cambiar la base de los demás valores del INPC, simplemente se les divide a todos entre ese mismo valor de 2000, 59.96995, con lo que se obtiene la serie que se muestra en la tabla 15.9.

Tabla 15.9 INPC con 2000 = 100.

Vigencia	INPC 2011 = 100	INPC 2000 = 100
01 ene. 2000	59.96995	100.00000
01 ene. 2001	64.68080	107.85535
01 ene. 2002	67.88722	113.20206
01 ene. 2003	71.26327	118.83163
01 ene. 2004	74.36092	123.99697
01 ene. 2005	77.60959	129.41413
01 ene. 2006	80.74517	134.64272
01 ene. 2007	83.96420	140.01046
01 ene. 2008	87.01013	145.08955
01 ene. 2009	92.40689	154.08866
01 ene. 2010	96.79132	161.39970
01 ene. 2011	100.282	167.22042

Ahora sí, para deflacionar la serie de los salarios mínimos, en la tabla 15.10 se reúnen los datos de los salarios mínimos de la zona A y los valores del INPC, con 2000 como año base.

Tabla 15.10 Datos para el ejemplo 15.11.

Vigencia	Salario mínimo Zona A	INPC 2000 = 100	Salario mínimo Zona A deflactado
01 ene. 2000	37.9	100.00000	37.90000
01 ene. 2001	40.35	107.85535	37.41122
01 ene. 2002	42.15	113.20206	37.23430
01 ene. 2003	43.65	118.83163	36.73264
01 ene. 2004	45.24	123.99697	36.48476
01 ene. 2005	46.8	129.41413	36.16298
01 ene. 2006	48.67	134.64272	36.14752
01 ene. 2007	50.57	140.01046	36.11873
01 ene. 2008	52.59	145.08955	36.24658

(*continúa*)

Tabla 15.10 (*continuación*)

Vigencia	Salario mínimo Zona A	INPC 2000 = 100	Salario mínimo Zona A deflactado
01 ene. 2009	54.8	154.08866	35.56394
01 ene. 2010	57.46	161.39970	35.60106
01 ene. 2011	59.82	167.22042	35.77314

Como se indicó, los valores deflactados de los salarios mínimos se obtienen dividiendo cada salario corriente (no deflactado) entre el correspondiente INPC y multiplicando el cociente por 100.

El primer valor deflactado no cambia, por supuesto, ya que se le divide y se le multiplica por 100. Para el año 2001, las operaciones son:

$$\frac{40.35}{107.85535}(100) = 37.41$$

Es claro que el hecho de que sea menor que 40.35 refleja la operación que se acaba de hacer: descontarle la inflación.

Observando el comportamiento de estos salarios deflactados se puede apreciar, entre otros detalles, que entre 2000 y 2011, el salario no sólo no aumentó en términos reales sino que disminuyó considerablemente: de $37.90 a $35.77, es decir, una disminución de (37.90 – 35.70) / 37.90 = 2.2 / 37.90 = 5.8%, lo cual a su vez refleja lo que la opinión pública da como hecho: que los salarios mínimos no compensan la pérdida de poder adquisitivo ocasionada por la inflación.

En la figura 15.2 se muestran las gráficas de estos salarios, nominales y deflactados, en la que se aprecia claramente cómo es que los salarios, una vez descontada la inflación, no subieron realmente sino que, más bien, vieron disminuir su poder adquisitivo.

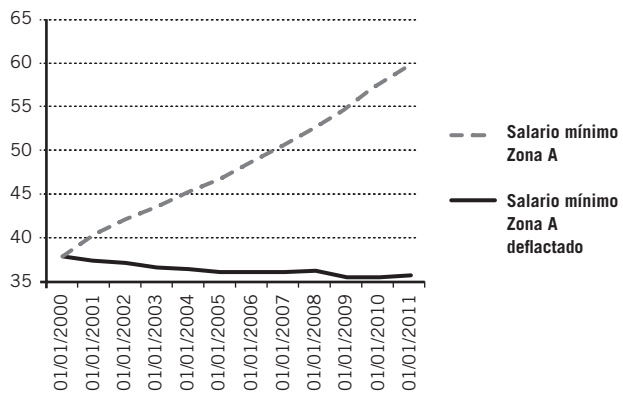


Figura 15.2 Salarios mínimos nominales y deflactados de la zona A.

15.6.1.5 El Índice Nacional de Precios al Productor

Tal como se anotó antes, el INPP es un índice que da seguimiento al comportamiento de los precios que los productores cobran a quienes les compran en primera instancia.

En realidad, el INPP es una familia de índices, ya que presenta resultados independientes para los 2 conjuntos de bienes y servicios en los que se divide la producción nacional: bienes finales y bienes intermedios. El conjunto de los primeros se constituye de bienes procesados como prendas de vestir, computadoras y automóviles; el segundo, de los bienes intermedios como productos agrícolas (maíz, algodón), partes de computadoras y partes automotrices. Las estimaciones de ambos conceptos se publican por separado. Los principales componentes del INPP se agrupan de acuerdo con 2 criterios: por el lado de la demanda (por destino o por sus consumidores) y por el lado de la oferta (por origen o por productores) de los bienes y servicios incluidos en la canasta.

Se puede encontrar una descripción detallada de la forma en la que se calcula este INPP en http://www.inegi.org.mx/est/contenidos/proyectos/inp/Presentacion.aspx y en la sección de *Estadísticas/Inflación y precios* del INEGI se pueden descargar las series históricas correspondientes.

15.6.2 Índices bursátiles

Los índices bursátiles se utilizan ampliamente para medir el comportamiento de los precios de las acciones que se cotizan en las diferentes bolsas de valores. En México, el índice más popular es el *Índice de Precios y Cotizaciones* de la Bolsa Mexicana de Valores (BMV), aunque la propia BMV calcula otros diversos índices: una visita sencilla a su sitio www.bmv.com.mx da una idea fácil de apreciar de los otros índices que se utilizan (en la sección de "Mercados de capitales", subsección "Resumen de índices"):

Índice	Último	Hora	Anterior	Variación %	Variación Ptos.	Máximo	Mínimo
IPC	37 241.29	09:17	37 242.05	0.00	–0.76	37 245.58	37 039.86
IMeBz	378.44	09:17	381.13	–0.71	–2.69	381.13	374.17
INMEX	2 160.94	09:17	2 168.37	–0.34	–7.43	2 168.37	2 152.25
IPC CompMx	286.36	09:17	286.44	–0.03	–0.08	286.47	284.87
IMC30	481.64	09:17	481.00	0.13	0.64	482.19	478.92
HABITA	502.52	09:16	494.30	1.66	8.22	503.23	494.15

Índice	Último	Hora	Anterior	Variación %	Variación Ptos.	Máximo	Mínimo
IPC SmallCap	205.95	09:15	205.02	0.45	0.93	206.02	204.97
IPC MidCap	186.78	09:17	187.11	–0.18	–0.33	187.69	185.86
IPC LargeCap	307.00	09:17	307.18	–0.06	–0.18	307.18	305.19
ÍNDICE DE DIVIDENDOS	154.52	09:17	155.21	–0.44	–0.69	155.27	154.06
BMV_Brasil 15							

Fuente: www.bmv.com.mx, consultado el 22 de febrero de 2011.

El Índice de Precios y Cotizaciones es el principal indicador de la Bolsa Mexicana de Valores, el cual expresa el rendimiento del mercado accionario en función de las variaciones de precios de una muestra balanceada, ponderada y representativa del total de acciones cotizadas en la Bolsa.

El IPC constituye un fiel indicador de las fluctuaciones del mercado accionario, gracias a 2 conceptos: por un lado, representatividad de la muestra en cuanto a la operatividad del mercado, que es asegurada mediante la selección de las emisoras líderes, determinadas a través de su nivel de bursatilidad y, por otra parte, la estructura de cálculo que contempla la dinámica del valor de capitalización de las emisoras que constituyen la muestra del IPC.

La muestra se conforma de 35 series accionarias, las cuales podrían variar en función de los criterios de selección o movimientos corporativos.

Por su parte, se calculan numerosos índices de precios para las acciones que cotizan en la Bolsa de Valores de Nueva York (NYSE, por sus siglas en inglés) que es la más importante del mundo. Al momento de escribir estas líneas, se llama NYSE–Euronext porque, tiempo atrás, se fusionó la NYSE original, con la bolsa Euronext de Europa. Ya desde antes de esa fusión la NYSE era la bolsa más importante del mundo y con la fusión adquirió mayor tamaño e importancia, además, hacia el momento de escribir este texto se anunció que la NYSE–Euronext se fusionaría con la Bolsa de Valores de Frankfurt, otra importante bolsa europea con lo que el tamaño y la importancia de esta megabolsa crecerá aún más.

Son muy numerosos los índices que se calculan para esta bolsa pero es posible que el más utilizado y, de hecho el más antiguo, sea el índice industrial Dow Jones (DJIA, por sus siglas en inglés); también, una búsqueda sencilla de *indexes* en el buscador de la NYSE arroja información sobre los diversos índices de precios que se calculan para esta bolsa:

Índices	Valor	Cambio (en $)	Cambio (en %)
Amex Composite	2 293.01	+17.13	0.75
Dow Jones Industrials	12 226.64	–41.55	0.34
Dow Jones Trans.	5 231.04	+13.14	0.25
Dow Jones Utilities	410.71	+0.95	0.23
Nasdaq Composite	2 804.35	–12.83	0.46
S & P 100	596.46	–1.80	0.30
S & P 400 Midcap	969.73	–4.65	0.48
S & P 500	1 328.01	–4.31	0.32
S & P Global 100	1 304.36	–0.21	0.02

Fuente: http://www.nyse.com/about/listed/mkt_indexes_other.shtml/, 19 de febrero de 2012.

EJEMPLO 15.13

En la tabla 15.11 se muestra una lista de 34 bolsas de valores del mundo, junto con el nombre del índice que las representa y la clave mediante la cual Yahoo! permite acceder a sus datos históricos.

Tabla 15.11 Lista de clave de Yahoo! para 34 índices de bolsa de valores.

	País	Índice	Clave Yahoo!
1	Alemania	DAX	^GDAXI
2	Argentina	MerVal	^MERV

(*continúa*)

Tabla 15.11 (*continuación*)

	País	Índice	Clave Yahoo!
3	Australia	All Ordinaries	^AORD
4	Austria	ATX	^ATX
5	Bélgica	BEL-20	^BFX
6	Brasil	Bovespa	^BVSP
7	Canadá	S&P TSX Composite	^GSPTSE
8	Chile	IPSA	^IPSA
9	China	Shanghai Composite	000001.SS
10	Corea del Sur	Seoul Composite	^KS11
11	Dinamarca	OMX Copenhagen 20	OMXC20.CO
12	Egipto	CMA	^CCSI
13	España	Madrid General	^SMSI
14	Estados Unidos	S&P 500	^GSPC
15	Filipinas	PSE Composite	^PSI
16	Francia	CAC 40	^FCHI
17	Holanda	AEX General	^AEX
18	Hong Kong	Hang Seng	^HSI
19	India	BSE 30	^BSESN
20	Indonesia	Jakarta Composite	^JKSE
21	Israel	TA-100	^TA100
22	Italia	MIBTel	^MIBTEL
23	Japón	Nikkei 225	^N225
24	Malasia	KLSE Composite	^KLSE
25	México	IPC	^MXX
26	Noruega	OSE All Share	^OSEAX
27	Nueva Zelanda	NZSE 50	^NZ50
28	Paquistán	Karachi 100	^KSE
29	Reino Unido	FTSE 100	^FTSE
30	Singapur	Straits Times	^STI
31	Sri Lanka	All Share	^CSE
32	Suiza	Swiss Market	^SSMI
33	Taiwán	Taiwan Weighted	^TWII
34	Turquía	ISE National-100	^XU100

Fuente: http://quote.yahoo.com/m2?u

a) Obtén los datos del último mes del índice Nikkei 225 de la Bolsa de Valores de Tokio, Japón, y grafica los precios de cierre.

b) Con los datos obtenidos en *a*), calcula los índices en cadena para días consecutivos y grafícalos.

Solución: *a*) En primer lugar, se accede a la sección de Finanzas del sitio de Yahoo!: http://mx.finance.yahoo.com/.

Del lado izquierdo, arriba de la gráfica del índice, aparece un espacio en blanco con una anotación a la derecha en la cual se lee "buscar cotizaciones".

Si se anota aquí ^N225 , y se le da clic, se llega a la página que contiene los datos de este índice y, del lado izquierdo, aparece una liga a "Precios históricos". Si ahora se da clic allí se accede a un listado de valores históricos del Nikkei 225, que es un listado con la fecha y diversos valores del índice: *apertura, máximo, mínimo, cierre, volumen y cierre ajustado*. De estos datos se separan los del cierre del último mes respecto a los que se tenían cuando se elaboró este ejemplo, que son los siguientes:

14 sep. 2011	8 518.57
15 sep. 2011	8 668.86
16 sep. 2011	8 864.16
20 sep. 2011	8 721.24
21 sep. 2011	8 741.16
22 sep. 2011	8 560.26
26 sep. 2011	8 374.13
27 sep. 2011	8 609.95
28 sep. 2011	8 615.65
29 sep. 2011	8 701.23
30 sep. 2011	8 700.29
03 oct. 2011	8 545.48
04 oct. 2011	8 456.12
05 oct. 2011	8 382.98
06 oct. 2011	8 522.02
07 oct. 2011	8 605.62
11 oct. 2011	8 773.68
12 oct. 2011	8 738.90
13 oct. 2011	8 823.25

Se grafican estos datos en la figura 15.4.

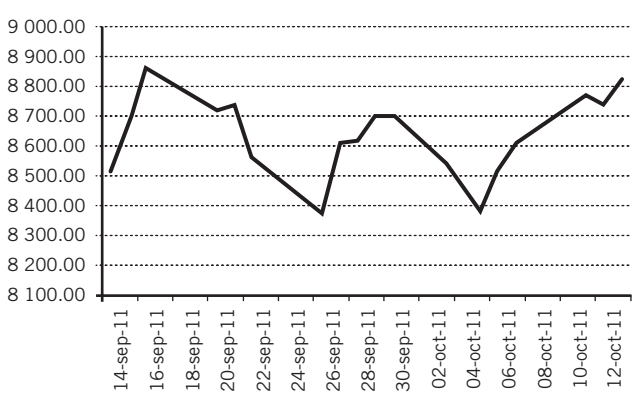


Figura 15.3 Índice Nikkei 225 del 14 de septiembre al 13 de octubre de 2011.

b) Con los datos obtenidos en *a*), calcule los índices en cadena para días consecutivos y grafíquelos.

En la tabla siguiente se muestran los índices en cadena y en la figura 15.4 se les grafica.

14 sep. 2011	
15 sep. 2011	101.7643
16 sep. 2011	102.2529
20 sep. 2011	98.38766
21 sep. 2011	100.2284
22 sep. 2011	97.93048
26 sep. 2011	97.82565
27 sep. 2011	102.8161
28 sep. 2011	100.0662
29 sep. 2011	100.9933
30 sep. 2011	99.9892
03 oct. 2011	98.22063
04 oct. 2011	98.9543
05 oct. 2011	99.13506
06 oct. 2011	101.6586
07 oct. 2011	100.981
11 oct. 2011	101.9529
12 oct. 2011	99.60359
13 oct. 2011	100.9652

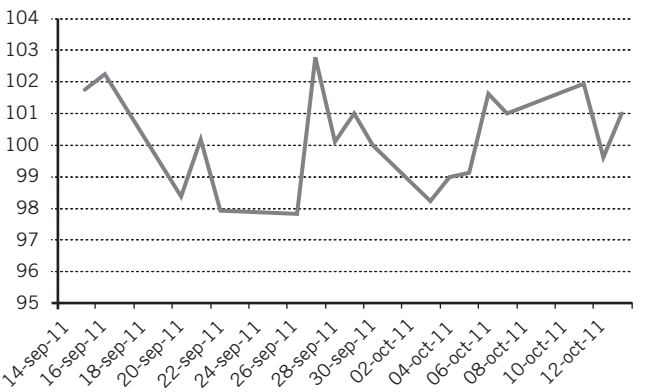


Figura 15.4 Índices en cadena del índice Nikkei 225 de la Bolsa de Valores de Tokio del 14 de septiembre al 13 de octubre de 2011.

ejercicios 15.6 Índices para propósitos especiales

1. Explique qué es el Índice Nacional de Precios al Consumidor.
2. ¿Qué es el Índice Nacional de Precios al Productor?
3. Enumere y explique cuáles son los principales usos del Índice Nacional de Precios al Consumidor.
4. Indique y explique cuáles son los principales usos del Índice Nacional de Precios al Productor.

Cálculo de la inflación mediante el INPC y cambio de periodo base

5. En la tabla siguiente se reunieron los índices de precios al consumidor mensuales de Estados Unidos de América para los meses de enero de 2008 a enero de 2011 con los datos del INPC de México para los mismos meses.

   a) Cambie el periodo base según se requiera para hacer que ambas series tengan el mismo periodo base.

   b) Calcule índices de inflación mensuales para ambas series, compárelos y comente.

Mes	INPC enero 2008 = 100	INPC EUA*
Ene. 2008	100.00000	211.08
Feb. 2008	100.42084	211.693
Mar. 2008	101.09926	213.528
Abr. 2008	101.30136	214.823
Mayo 2008	101.20784	216.632
Jun. 2008	101.58271	218.815
Jul. 2008	102.23894	219.964
Ago. 2008	102.83572	219.086
Sep. 2008	103.50304	218.783
Oct. 2008	104.19018	216.573
Nov. 2008	105.39088	212.425
Dic. 2008	106.15331	210.228
Ene. 2009	106.20245	211
Feb. 2009	106.59951	212.193
Mar. 2009	107.27951	212.709
Abr. 2009	107.52600	213.24
Mayo 2009	107.17569	213.856
Jun. 2009	107.40711	215.693
Jul. 2009	107.68847	215.351
Ago. 2009	108.01975	215.834
Sep. 2009	108.46437	215.969
Oct. 2009	108.72591	216.177
Nov. 2009	109.39798	216.33
Dic. 2009	109.92265	215.949
Ene. 2010	111.24144	216.687
Feb. 2010	111.82634	216.741
Mar. 2010	112.51823	217.631
Abr. 2010	111.96741	218.009

(*continúa*)

(*continuación*)

Mes	INPC enero 2008 = 100	INPC EUA*
Mayo 2010	111.36111	218.178
Jun. 2010	111.34447	217.965
Jul. 2010	111.63216	218.011
Ago. 2010	111.95156	218.312
Sep. 2010	112.53091	218.439
Oct. 2010	113.24816	218.711
Nov. 2010	114.11600	218.803
Dic. 2010	114.92915	219.179
Ene. 2011	115.25325	220.223

*ftp://ftp.bls.gov/pub/special.requests/cpi/cpiai.txt, 1982-84=100, consultado el 20 de febrero de 2012.

Deflación de series de tiempo

6. Con los datos de la tabla siguiente de salarios mínimos de las zonas B y C, y del INPC:

Año	Salario mínimo zona B	Salario mínimo zona C	INPC 2000 = 100 enero
2004	43.73	42.11	124.133734464700
2005	45.35	44.05	129.775390800196
2006	47.16	45.81	134.882052543483
2007	49.00	47.60	140.251599560528
2008	50.96	49.50	145.447042733988
2009	53.26	51.95	154.584612008217
2010	55.84	54.47	161.474973394059
2011	58.13	56.70	167.582016959414

a) Grafique los salarios mínimos para la zona B.
b) Deflacte la serie de tiempo.
c) Grafique los salarios reales para la zona B.

7. Con los datos de la tabla del ejercicio 6:

a) Grafique los salarios mínimos para la zona C.
b) Deflacte la serie de tiempo.
c) Grafique los salarios reales para la zona C.

8. En la siguiente tabla se muestra el comportamiento del salario mínimo para un maestro de escuela primaria particular en la zona A:

a) Grafique los salarios mínimos.
b) Deflacte la serie de tiempo.
c) Grafique los salarios reales.

Año	Salario mínimo	INPC 2000 = 100 enero
2004	69.70	124.133734464700
2005	72.10	129.775390800196
2006	74.98	134.882052543483
2007	77.90	140.251599560528
2008	81.02	145.447042733988
2009	84.42	154.584612008217
2010	88.51	161.474973394059
2011	92.14	167.582016959414

9. En la siguiente tabla se muestra la evolución de los salarios mínimos para un reportero gráfico en prensa diaria impresa en la zona A:

a) Grafique los salarios mínimos.
b) Deflacte la serie de tiempo.
c) Grafique los salarios reales.

Año	Salario mínimo	INPC 2000 = 100 enero
2004	135.51	124.133734464700
2005	140.20	129.775390800196
2006	145.81	134.882052543483
2007	151.50	140.251599560528
2008	157.56	145.447042733988
2009	164.18	154.584612008217
2010	172.14	161.474973394059
2011	179.20	167.582016959414

10. En la siguiente tabla se muestran los salarios mínimos para un cajero de máquina registradora en la zona A:

a) Grafique los salarios mínimos.
b) Deflacte la serie de tiempo.
c) Grafique los salarios reales.

Año	Salario mínimo	INPC 2000 = 100 enero
2004	58.50	124.133734464700
2005	60.50	129.775390800196
2006	62.92	134.882052543483
2007	65.37	140.251599560528
2008	67.98	145.447042733988
2009	70.84	154.584612008217
2010	74.28	161.474973394059
2011	77.33	167.582016959414

Índices bursátiles

11. Encuentre los valores mensuales del último año del Índice de Precios y Cotizaciones de la Bolsa Mexicana de

Valores y los del Índice Industrial Dow Jones de la Bolsa de Valores de Nueva York.

a) Determine los incrementos porcentuales mensuales de los 2 índices.
b) Compare los aumentos mensuales y comente.

12. Utilice las claves de Yahoo! para los datos de 34 índices de bolsas de valores del mundo de la tabla 15.13.

a) Obtenga los datos del último mes del índice (escoger 1 o varios) y grafique los precios de cierre.
b) Con los datos que obtuvo, calcule índices en cadena para días consecutivos y grafíquelos.

15.7 Números índices y Excel

Excel no cuenta con funciones específicas para calcular números índices; sin embargo, el uso de su hoja de cálculo, con la facilidad que permite hacer cálculos repetidos con copiar y pegar, es de gran ayuda con, por ejemplo, el cálculo de los índices en cadena tal como se menciona en la sección 15.5.

Es claro que el estudiante sabrá aprovechar las potencialidades ordinarias de este paquete para ayudarse en sus labores.

15.8 Resumen

Se revisó aquí el tema de los números índice, que son valores relativos, cocientes, que comparan el precio, la cantidad o el valor de uno o varios artículos de un periodo dado contra el precio, la cantidad o el valor de un periodo base.

Se vio que los índices simples son comparaciones de un precio, una cantidad o un valor para un solo artículo, en tanto que los precios agregados son comparaciones de varios precios, varias cantidades o varios valores.

Se revisaron 3 índices agregados de valor propuestos por otros tantos estadísticos: de Laspeyres, de Paasche y el ideal de Fischer.

Se vio también que los principales usos de los números índice son:

- Comparar datos dados en diferentes unidades, ya que los números índice son, precisamente, números relativos expresados en porcentaje.
- Comparar conjuntos de artículos, a través de los índices agregados.
- Eliminar en series de datos los efectos de la inflación, es decir, *deflacionar*.

Finalmente se vieron 2 grupos de índices para propósitos especiales que se utilizan ampliamente: índices de precios al consumidor y al productor e índices bursátiles. De la primera categoría, se vieron el Índice Nacional de Precios al Consumidor y el Índice Nacional de Precios al Productor; para el segundo rubro se revisaron el Índice de Precios y Cotizaciones de la Bolsa Mexicana de Valores y el Índice Industrial Dow Jones de la Bolsa de Valores de Nueva York.

Por último, con datos del INPC se ilustraron los procedimientos para cambiar la base de una serie de índices y para deflacionar series.

15.9 Fórmulas del capítulo

15.2 Números índice simples

Índice simple de precios:

$$ISP = \frac{P_n}{P_0}(100) \qquad \textbf{(15.1)}$$

Índice simple de cantidades:

$$ISQ = \frac{Q_n}{Q_0}(100) \qquad \textbf{(15.2)}$$

Índice simple de valor:

$$ISV = \frac{P_n Q_n}{P_0 Q_0}(100) \qquad \textbf{(15.3)}$$

15.3 Números índice agregados

Índice agregado de precios:

$$IAP = \frac{\sum P_n}{\sum P_0}(100) \qquad \textbf{(15.4)}$$

Índice agregado de cantidad:

$$IAQ = \frac{\sum Q_n}{\sum Q_0}(100) \qquad \textbf{(15.5)}$$

Índice agregado de valor:

$$IAV = \frac{\sum P_n Q_n}{\sum P_0 Q_0}(100) \qquad \textbf{(15.6)}$$

15.4.1 Índice de Laspeyres

Índice de valores de Laspeyres:

$$IVL = \frac{\sum P_n Q_0}{\sum P_0 Q_0}(100) \qquad \textbf{(15.7)}$$

15.4.3 Índice ideal de Fischer

Índice de valor ideal de Fischer:

$$IVF = \sqrt{\left(\frac{\sum P_n Q_0}{\sum P_0 Q_0}\right)\left(\frac{\sum P_n Q_n}{\sum P_0 Q_n}\right)}(100) \qquad \textbf{(15.9)}$$

15.4. Índice de Paasche

Índice de valores de Paasche:

$$IVP = \frac{\sum P_n Q_n}{\sum P_0 Q_n}(100) \qquad \textbf{(15.8)}$$

15.10 Ejercicios adicionales

15.2 a 15.5 Números índice simples, agregados y en cadena

1. En la tabla siguiente se muestran las colegiaturas y el número de alumnos que tiene una escuela primaria particular, en 2 ciclos escolares, 2011 y 2012:

	Ciclo 2011		Ciclo 2012	
Grado	Colegiatura P_0	Núm. de alumnos Q_0	Colegiatura P_n	Núm. de alumnos Q_n
Primero	1 200	125	1 800	140
Segundo	1 575	75	2 250	60
Tercero	2 250	240	3 000	240
Cuarto	2 625	220	3 450	210
Quinto	3 750	205	4 500	230
Sexto	4 350	210	5 250	225

a) Calcule el índice agregado de precios (colegiaturas) para el ciclo 2012 respecto al ciclo 2011.
b) Calcule el índice agregado de cantidades (número de alumnos) para el ciclo 2012 respecto al ciclo 2011.
c) Calcule el índice agregado del valor total de las colegiaturas para el ciclo 2012 respecto al ciclo 2011.
d) Calcule los índices de Laspeyres, Paasche y Fischer para el ciclo 2012 respecto al ciclo 2011.

2. Para los datos siguientes sobre producción de cereales:

a) Calcule índices simples de precios para cada producto sucesivamente para cada par de años y comente.
b) Calcule índices agregados simples comparando sucesivamente cada par de años (índices en cadena) y comente.

	Cebada	Maíz	Arroz	Sorgo	Trigo
	USD por tonelada métrica				
Ene. 2007	152.13	165.1	313.48	175.05	196.07
Ene. 2008	205.61	206.53	393.48	212.67	369.59
Ene. 2009	121.61	173.24	615.25	153.26	239.36
Ene. 2010	146.6	167.21	598	161.79	201.51
Ene. 2011	195.12	265.29	528.38	246.32	326.55

3. Una empresa desea un análisis de los cambios en los precios de 3 de sus productos en los 3 últimos años. Los datos con los que cuenta son los siguientes:

	Precios			Cantidades		
Producto	2010 P_0	2011 P_n	2012 P_n	2010 Q_0	2011 Q_n	2012 Q_n
A	$ 100	$ 55	$ 200	1 500	1 700	1 600
B	$ 30	$ 50	$ 75	550	680	1 200
C	$ 90	$ 50	$ 750	1 000	900	850

a) Calcule los índices de Laspeyres y de Paasche, con 2010 = 100.
b) Construya el índice ideal de Fischer.

4. El periodo de 1980 a 1990 fue de hiperinflación en México. En la tabla siguiente se muestran los precios nominales al público de la gasolina que entonces se llamaba *Nova* para varios años consecutivos.

Año	Precio
1981	$6
1982	$20
1983	$30
1984	$40
1985	$80
1986	$155
1987	$367.5

a) Calcule el índice de precios simple para la gasolina, con 1981 = 100 como año base.
b) Calcule el índice simple para 1987 con 1981 = 100.
c) Calcule el precio de la gasolina para 1988 si se sabe que el índice con base 1981 = 100, fue de 8 000.

5. Durante una negociación para evaluar un aumento en las tarifas del transporte urbano, los propietarios de los vehículos solicitaban un aumento en las tarifas, de $2.00 a $3.50; para apoyar su petición, publicaron la siguiente información sobre los aumentos en sus costos.

	Gastos (pesos)		
Concepto	2006 P_0	2011 P_n	Índices simples
Cambio de aceite	320	500	156.25
Llantas	2 000	2 500	125.00
Afinación	400	500	125.00
Cuotas	350	500	142.86
			549.11

$$IAP = \frac{\sum P_n}{\sum P_0}(100) = \frac{4\ 000}{3\ 070}(100) = 130.29$$

¿Se justifica el reclamo de los propietarios de los vehículos de transporte urbano?

15.6 Índices para propósitos especiales

6. Considere los siguientes valores del INPC para los siguientes salarios. Convierta estos salarios, que están en pesos corrientes, a pesos constantes de 2008.

Año	Salario	INPC	Salario en pesos constantes de 2008
2008	10 000	100.0	10 000
2009	10 500	104.3	10 067.11
2010	10 800	109.9	9 827.12
2011	11 300	116.4	9 707.08

7. En la tabla siguiente se muestran datos del INPC y del Producto Interno Bruto (flujos corrientes octubre-diciembre, a precios de mercado).

Año	PIB	INPC
2003	7 293 558.2	73.783729734576
2004	7 936 731.6	77.613731182722
2005	9 062 193.6	80.200395826581
2006	9 712 554.1	83.451138863412
2007	10 776 388.1	86.588098998021
2008	11 942 415.5	92.240695661768
2009	12 218 270.5	95.536951859488
2010	12 623 583.1	99.742092088296

Considerando esta información, ¿hubo aumento en los ingresos reales de esos años?

8. Utilizando las claves de Yahoo! para los datos de 34 índices de bolsas de valores del mundo de la tabla 15.13:

a) Obtenga los datos del último mes del índice (escoger uno o varios) y grafique los precios de cierre.
b) Con los datos obtenidos en *a*), calcule los índices en cadena para días consecutivos y grafíquelos.

capítulo 16

Análisis de series de tiempo

Sumario

16.1 Modelo clásico de series de tiempo
16.2 Análisis gráfico de la tendencia
16.3 Tendencia secular
- 16.3.1 Suavización con promedios móviles exponenciales
- 16.3.2 Ajuste de una recta con mínimos cuadrados
- 16.3.3 Ajuste de una función exponencial con mínimos cuadrados
- 16.3.4 Ajuste de una parábola con mínimos cuadrados

16.4 Variaciones estacionales
- 16.4 1 Cálculo de índices estacionales
- 16.4.2 Desestacionalización de series de tiempo
- 16.4.3 Pronósticos con índices estacionales

16.5 Variaciones cíclicas
16.6 Resumen
16.7 Fórmulas del capítulo
16.8 Ejercicios adicionales

Serie de tiempo. Conjunto de observaciones de alguna variable tomadas a intervalos regulares.

Una **serie de tiempo** es un conjunto de observaciones de alguna variable, tomadas a intervalos regulares. Algunos ejemplos de series de tiempo son los datos diarios de los precios de alguna acción que cotiza en la bolsa de valores, o los datos mensuales de la inflación o los datos anuales de ventas de una empresa.

El análisis de series de tiempo tiene como propósitos principales, por un lado, estudiar las variaciones que se dan en los valores de la serie a lo largo del tiempo y, por otra parte, hacer pronósticos sobre su posible comportamiento futuro.

16.1 Modelo clásico de series de tiempo

Tradicionalmente se considera que las series de tiempo se forman de 4 componentes: tendencia secular (T), variaciones estacionales (E), variaciones cíclicas (C) y variaciones irregulares (I). El conjunto de estos 4 componentes da lugar a los valores que se observan en la serie de tiempo.

La *tendencia secular* en una serie de tiempo representa el movimiento básico a largo plazo de la serie, mientras que las *variaciones cíclicas* se manifiestan como variaciones por encima y por debajo de la tendencia y tienen una duración de más de un año. Por su parte, las *variaciones estacionales* son movimientos que se dan a plazo de cuando mucho un año como, por ejemplo, los aumentos en las ventas de juguetes hacia cada fin de año o el aumento de los artículos para vacacionar que ocurre en el verano. Por último, las *variaciones irregulares* son aquellas que suceden en forma aleatoria respecto a la tendencia y que no se atribuyen a variaciones cíclicas ni estacionales.

La interpretación tradicional de estos 4 componentes se da en 2 formas, conocidas como el *modelo multiplicativo* y el *modelo aditivo*. Si se utiliza Y para identificar el valor observado de la variable, que es la variable que depende de los 4 componentes, entonces el modelo aditivo quedaría expresado como:

$$Y = T + C + E + I \tag{16.1}$$

En este modelo, las variaciones estacionales, cíclicas e irregulares son desviaciones cuantitativas respecto a la tendencia secular y se asume que los componentes son independientes entre sí. Por su parte, el modelo multiplicativo, que es el que se utiliza más comúnmente, se representaría simbólicamente como:

$$Y = T \cdot E \cdot C \cdot I \tag{16.2}$$

De los 4 componentes mencionados, el más importante es el de la tendencia secular, ya que contiene la tendencia principal de la serie de tiempo, la de largo plazo. En la tabla 16.1 se muestran los datos del Índice Nacional de Precios al Consumidor (INPC), de julio de 2009 a junio de 2011, y en la figura 16.1 se muestra su gráfica.

Tabla 16.1 Datos del INPC (julio de 2009 a junio de 2011)

Jul. 2009	93.671601856385	Jul. 2010	97.077503396247
Ago. 2009	93.895719694096	Ago. 2010	97.347134394847
Sep. 2009	94.366711949963	Sep. 2010	97.857433471482
Oct. 2009	94.652203595540	Oct. 2010	98.461517243282
Nov. 2009	95.143194058464	Nov. 2010	99.250412032025
Dic. 2009	95.536951859488	Dic. 2010	99.742092088296
Ene. 2010	96.575479439774	Ene. 2011	100.228000000000
Feb. 2010	97.134050050685	Feb. 2011	100.604000000000
Mar. 2010	97.823643397489	Mar. 2011	100.797000000000
Abr. 2010	97.511947204733	Abr. 2011	100.789000000000
Mayo 2010	96.897519532732	Mayo 2011	100.046000000000
Jun. 2010	96.867177425472	Jun. 2011	100.041000000000

Esta gráfica muestra el comportamiento de largo plazo de este indicador económico que, se sabe, es afectado por numerosos factores. La gráfica muestra el comportamiento ascendente de este INPC y el propósito del análisis de las series de tiempo es identificar esta tendencia de largo plazo que pudiera ser aproximadamente en línea recta, o lineal, como en esa figura, pero que también puede tener comportamientos más semejantes a curvas, simples o complejas, que a rectas.

En la figura 16.1 parecen también apreciarse variaciones estacionales en las 2 gibas que se ven en la línea de tendencia, aproximadamente en marzo de 2010 y marzo de 2011.

Aunque esta gráfica contiene datos de solamente 24 meses, 2 años, en series de tiempo más largas también se consideran las variaciones cíclicas cuando se trata de comprender el comportamiento de una serie de tiempo o cuando se trata de pronosticar su comportamiento futuro.

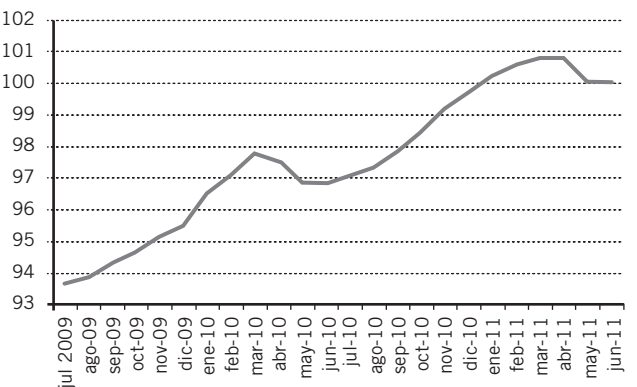


Figura 16.1 Gráfica de los datos del INPC (julio de 2009 a junio de 2011).

El análisis de las series de tiempo se hace de varias maneras pero con la idea básica de eliminar algunos componentes para poder apreciar otros. Así, teóricamente, si se eliminan de una serie de tiempo los componentes de tendencia y las variaciones estacionales, lo que queda es el conjunto de las variaciones cíclicas y las irregulares. Estas últimas son las que se producen por 2 principales razones: en primer lugar, por sucesos extraordinarios o imprevisibles como elecciones, temblores, inundaciones, cracs bursátiles, etc. El otro tipo de variaciones irregulares se deben a eventos enteramente aleatorios e impredecibles que, por su misma naturaleza, no son identificables.

Si se eliminan, finalmente, estas variaciones irregulares, lo que queda en la serie de tiempo es su componente cíclico. En teoría, este solo componente aislado muestra la variación cíclica de los negocios o de la variable que se mide y las diferencias fundamentales en relación con las variaciones estacionales, pues su duración es superior a un año y las causas que las originan son enteramente diferentes. El muy estudiado ciclo de negocios, que se compone de periodos sucesivos de prosperidad, recesión, depresión y recuperación es el resultado de variables que no tienen que ver con el clima, las costumbres sociales y las otras variables que dan cuenta de las variaciones estacionales. El estudio de los ciclos es un tema de gran importancia en sí mismo y da lugar a una gran cantidad de estudios especializados. En 1939, Schumpeter[1] compiló la siguiente lista de ciclos: de Kitchin, con duración de 3 años; de Juglar, con duración de entre 9 y 10 años; de Kuznetz, con duración de entre 15 y 20 años; y los de Kondratiev, con duración de entre 48 y 60 años.

[1] Schumpeter, Joseph A., *Business Cycles: A theoretical, historical and statistical analysis of the Capitalist process*, McGraw-Hill, Nueva York, 1939.

En las secciones siguientes se revisan los métodos que más comúnmente se utilizan para analizar la tendencia secular y las variaciones estacionales y cíclicas pero comenzando por el breve pero muy importante análisis gráfico de la serie.

16.2 Análisis gráfico de la tendencia

Al igual que en el análisis de regresión, un paso inicial muy útil para analizar series de tiempo consiste en graficar los datos en un diagrama de dispersión. En este tipo de gráficas para series de tiempo, el tiempo es la variable independiente y la dependiente es la de interés, como el INPC en la figura 16.1.

Observar los puntos en la gráfica puede dar una buena idea sobre si los datos muestran una tendencia clara o no; en caso de existir se puede identificar si ésta podría quedar bien descrita mediante una línea recta o si, más bien, su comportamiento se asemeja a algún tipo de curva y puede, por lo tanto, sugerir la presencia de alguno de los otros componentes: variaciones cíclicas, estacionales o irregulares. Esta revisión preliminar es muy útil como auxiliar para determinar el modelo que debe emplearse para el análisis.

EJEMPLO 16.1

En la tabla 16.2 se presentan los datos de cierre de año del Índice de Precios y Cotizaciones (IPC) de la Bolsa Mexicana de Valores (BMV), de 1985 a 2010, y en la figura 16.2 se muestra la gráfica correspondiente.

Tabla 16.2 Cierres de año del IPC de la BMV

1985	11.082
1986	47.101
1987	105.670
1988	211.532
1989	418.925
1990	628.790
1991	1 431.460
1992	1 759.440
1993	2 602.630
1994	2 375.660
1995	2 778.470
1996	3 361.030
1997	5 229.350
1998	3 959.660
1999	7 129.880
2000	5 652.190
2001	6 372.280
2002	6 127.090
2003	8 795.280
2004	12 917.880
2005	17 802.710
2006	26 448.320
2007	29 536.830
2008	22 380.320
2009	32 120.470
2010	38 550.790

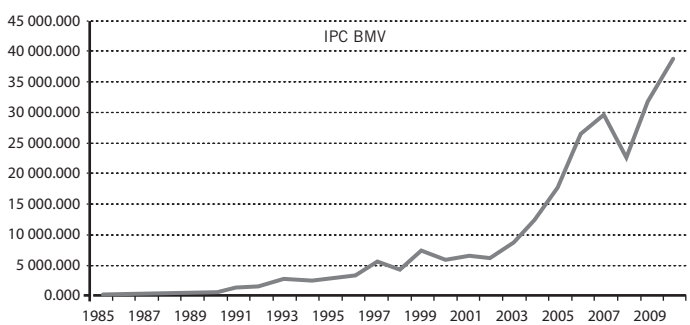


Figura 16.2 Gráfica de cierres de año del IPC de la BMV, 1985-2010.

Una primera aproximación visual muestra una tendencia ascendente curva, lo cual sería una muestra inicial de que el modelo que se adapta a una forma así no es lineal y que el adecuado podría ser un modelo curvo como el de una parábola.

Sin embargo, también puede apreciarse en la gráfica un quiebre en 2002, en donde terminaría una primera racha ascendente con una inclinación mucho menor que la arranca de ahí. Por ello, también podría evaluarse la posibilidad de partir la serie en ese año para utilizar los años 2002-2010 para hacer el análisis. Por supuesto, lo que se decida hacer depende de varios factores, entre los que destaca el propósito del estudio.

16.3 Tendencia secular

En el análisis de la tendencia secular se utilizan principalmente 2 métodos:

1. Método de suavización mediante promedios móviles (estudiado en la sección 3.1.5 del capítulo 3).
2. Método de ajuste de líneas a la tendencia mediante el método de mínimos cuadrados (analizado a detalle en el capítulo 13).

Se revisan los detalles de ambos en las subsecciones siguientes.

16.3.1 Suavización con promedios móviles exponenciales

Suele suceder que las series de tiempo presentan claras muestras de variaciones, por ello se utilizan los promedios móviles como los promedios móviles simples que se estudiaron en el capítulo 3, para eliminar o reducir fluctuaciones debidas a variaciones estacionales, cíclicas e irregulares. En la sección 3.1.5 de ese capítulo se analizaron los promedios móviles simples pero, en el caso de las series de tiempo, es más común que se aplique otro tipo de promedios móviles para realizar la suavización, que son los promedios móviles exponenciales.

Estos promedios móviles exponenciales se definen como:

$$PME_t = PME_{t-1} + w\left(Y_t - PME_{t-1}\right) \tag{16.3}$$

En donde:

PME_t es el promedio móvil exponencial del periodo.
PME_{t-1} es el promedio móvil exponencial del periodo anterior.
Y_t es la observación de la serie original correspondiente al periodo.
w es un factor de suavización, o peso, que asume valores entre 0 y 1.

Como puede verse, al revisar la fórmula, cada promedio móvil exponencial se construye sumando al PME del periodo anterior una porción (w) de la diferencia entre el valor observado de la serie en el periodo y el promedio móvil exponencial del anterior. Este procedimiento hace que, de alguna manera, cada promedio móvil subsecuente conserve algo de la "memoria" de lo que sucedió en el pasado, ya que todos ellos se construyen con datos previos.

Es importante la selección del factor de suavización, pues el uso de un valor cercano a 0 hace que se dé menor peso a los valores más recientes de la serie de tiempo, en tanto que los valores cercanos a 1 dan mayor peso a estos valores más recientes. En otras palabras, un valor de w cercano a 0 hace que la fracción de la diferencia entre el PME del día y el PME del periodo anterior que se suma sea mucho menor que cuando w es cercano a 1. Se aprecia esto en el ejemplo siguiente.

Ejemplo 16.2

En la tabla 16.3 se muestran nuevamente los datos de los cierres del IPC de la BMV vistos antes, y en la que se incluyeron los cálculos de promedios móviles exponenciales, con $w = 0.1$, con $w = 0.5$ y otro con $w = 0.9$. Recuerde que, como se necesita un valor inicial para un promedio móvil y no se tiene, se comienza con el primer valor de la serie, en este caso, con el valor de 11.082 del IPC para 1985.

Tabla 16.3 IPC con PME con $w = 0.1$, w = 0.5 y $w = 0.9$

Año	IPC	$w = 0.1$	$w = 0.5$	$w = 0.9$
1985	11.082			
1986	47.101	14.6839	29.0915	43.4991
1987	105.67	23.78251	67.38075	99.45291
1988	211.532	42.55746	139.4564	200.3241
1989	418.925	80.19421	279.1907	397.0649
1990	628.79	135.0538	453.9903	605.6175
1991	1 431.46	264.6944	942.7252	1 348.876
1992	1 759.44	414.169	1 351.083	1 718.384
1993	2 602.63	633.0151	1 976.856	2 514.205
1994	2 375.66	807.2796	2 176.258	2 389.515
1995	2 778.47	1 004.399	2 477.364	2 739.574
1996	3 361.03	1 240.062	2 919.197	3 298.884
1997	5 229.35	1 638.991	4 074.274	5 036.303
1998	3 959.66	1 871.058	4 016.967	4 067.324
1999	7 129.88	2 396.94	5 573.423	6 823.624
2000	5 652.19	2 722.465	5 612.807	5 769.333
2001	6 372.28	3 087.446	5 992.543	6 311.985
2002	6 127.09	3 391.411	6 059.817	6 145.58
2003	8 795.28	3 931.798	7 427.548	8 530.31
2004	12 917.88	4 830.406	10 172.71	12 479.12
2005	17 802.71	6 127.636	13 987.71	17 270.35
2006	26 448.32	8 159.705	20 218.02	25 530.52
2007	29 536.83	10 297.42	24 877.42	29 136.2
2008	22 380.32	11 505.71	23 628.87	23 055.91
2009	32 120.47	13 567.18	27 874.67	31 214.01
2010	38 550.79	16 065.54	33 212.73	37 817.11

Ahora, en la figura 16.3 se muestran las gráficas de los 3 PME junto con la gráfica de los datos originales del IPC.

Como puede verse en la figura 16.3, el PME con $w = 0.9$ es casi indistinguible del lPC (es muy grande la porción de la diferencia entre el IPC del periodo menos el PME del anterior, que se suma al PME anterior, 0.9 o 90%), en tanto que, con $w = 0.1$, se distinguen claramente la línea del IPC y la trayectoria del PME (es muy pequeña la porción de la diferencia entre el IPC del periodo actual menos el PME del periodo anterior, que se suma al PME anterior, 0.1 o 10%).

Por otra parte, la gráfica del IPC con un PME con $w = 0.5$, en la que puede apreciarse el efecto suavizador de los promedios móviles: las curvas que describe la línea punteada del PME son más suaves, con quiebres menos marcados que los datos originales del IPC.

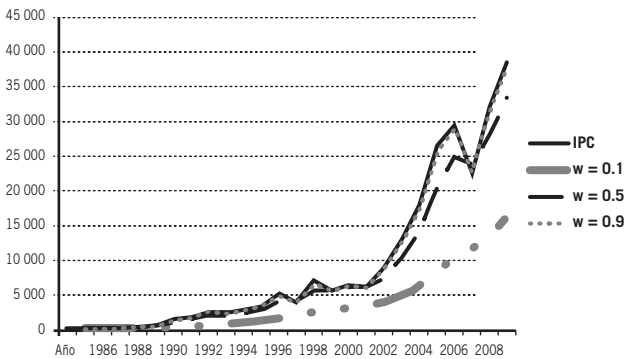


Figura 16.3 IPC con PME con $w = 0.1$, 0.5 y 0.9.

En la subsección siguiente se analiza el procedimiento para utilizar promedios móviles para hacer pronósticos sobre series de tiempo.

16.3.1.1 Pronósticos con promedios móviles exponenciales

Los pronósticos basados en promedios móviles no se hacen sobre los datos originales sino sobre los promedios móviles mismos y sólo son convenientes para pronosticar un periodo hacia adelante y en plazos cortos.

Los ejemplos que se vieron antes con datos anuales del IPC de la BMV se utilizaron para ilustrar las propiedades y el comportamiento de los promedios móviles exponenciales; ahora, para ilustrar el procedimiento mediante el cual se pueden hacer pronósticos basados en promedios móviles exponenciales, se utilizan datos diarios de cotizaciones bursátiles. El procedimiento para hacer esto consiste en:

1. Calcular los PME.
2. Calcular el pronóstico con la misma fórmula 16.3:

$$PME_t = PME_{t-1} + w\left(Y_t - PME_{t-1}\right) \tag{16.4}$$

Ejemplo 16.3

Con los datos de las 2 primeras columnas de la tabla 16.4, que son precios diarios de cierre de las acciones de Geo B:

a) Calcule promedios móviles exponenciales con $w = 0.1$, $w = 0.5$ y $w = 0.9$.

b) Haga pronósticos con los mismos valores de w para el 15 de julio de 2011.

Tabla 16.4 Precios de cierre de las acciones de Geo B y cálculos para el ejemplo 16.3

Fecha		$w = 0.1$	$w = 0.5$	$w = 0.9$
23 mayo 2011	28.25			
24 mayo 2011	27.40	28.165	27.825	27.485
25 mayo 2011	27.43	28.0915	27.6275	27.4355
26 mayo 2011	28.55	28.13735	28.08875	28.43855
27 mayo 2011	28.38	28.16162	28.23438	28.38586
30 mayo 2011	28.20	28.16545	28.21719	28.21859
31 mayo 2011	27.42	28.09091	27.81859	27.49986
01 jun. 2011	27.10	27.99182	27.4593	27.13999
02 jun. 2011	26.90	27.88264	27.17965	26.924
03 jun. 2011	26.51	27.74537	26.84482	26.5514
06 jun. 2011	26.21	27.59183	26.52741	26.24414
07 jun. 2011	26.00	27.43265	26.26371	26.02441
08 jun. 2011	26.05	27.29439	26.15685	26.04744
09 jun. 2011	26.30	27.19495	26.22843	26.27474
10 jun. 2011	25.92	27.06745	26.07421	25.95547
13 jun. 2011	25.81	26.94171	25.94211	25.82455
14 jun. 2011	25.80	26.82754	25.87105	25.80245
15 jun. 2011	25.68	26.71278	25.77553	25.69225
16 jun. 2011	25.60	26.6015	25.68776	25.60922
17 jun. 2011	25.40	26.48135	25.54388	25.42092

Fecha		$w = 0.1$	$w = 0.5$	$w = 0.9$
06 jun. 2011	25.05	26.33822	25.29694	25.08709
21 jun. 2011	25.08	26.2124	25.18847	25.08071
22 jun. 2011	25.53	26.14416	25.35924	25.48507
23 jun. 2011	25.74	26.10374	25.54962	25.71451
24 jun. 2011	26.39	26.13237	25.96981	26.32245
27 jun. 2011	26.50	26.16913	26.2349	26.48225
28 jun. 2011	26.79	26.23122	26.51245	26.75922
29 jun. 2011	26.99	26.3071	26.75123	26.96692
30 jun. 2011	27.00	26.37639	26.87561	26.99669
01 jul. 2011	27.45	26.48375	27.16281	27.40467
04 jul. 2011	27.65	26.60037	27.4064	27.62547
05 jul. 2011	27.06	26.64634	27.2332	27.11655
06 jul. 2011	26.36	26.6177	26.7966	26.43565
07 jul. 2011	26.75	26.63093	26.7733	26.71857
08 jul. 2011	26.74	26.64184	26.75665	26.73786
11 jul. 2011	25.67	26.54465	26.21333	25.77679
12 jul. 2011	25.68	26.45819	25.94666	25.68968
13 jul. 2011	26.15	26.42737	26.04833	26.10397
14 jul. 2011	25.48	26.33263	25.76417	25.5424
15 jul. 2011	24.65	26.16437	25.20708	24.73924

Solución: *a*) Los cálculos de los PME se incluyen en la tabla 16.4.

El verdadero valor de Geo B el 15 de julio de 2011 fue 24.65. Ahora, para analizar qué tan preciso es cada pronóstico respecto al verdadero valor de cierre, se resumen los pronósticos en el siguiente cuadro:

	Geo B 15 jul 2011	Pronóstico
$w = 0.1$	24.65	26.164
$w = 0.5$	24.65	25.207
$w = 0.9$	24.65	24.739

En este caso, el pronóstico más acertado fue el que se obtuvo con $w = 0.9$. Sin embargo, no siempre este valor del factor de suavización es el más preciso.

En libros especializados de pronósticos con series de tiempo se encuentran técnicas y refinamientos adicionales de estos mecanismos para realizar pronósticos con promedios móviles exponenciales.

16.3.2 Ajuste de una recta con mínimos cuadrados

En esta sección se repasa la forma en la que ajustan las rectas a una nube de puntos o diagrama de dispersión mediante el método de mínimos cuadrados; se dice "se repasa", porque este tema ya se abordó en el capítulo 13, que se ocupa del análisis de regresión y correlación lineal. Sin embargo, en esta sección se añaden 2 procedimientos que no se vieron en ese capítulo 13:

1. El ajuste de una parábola y
2. el ajuste de una función exponencial

Ambos por el método de mínimos cuadrados.

EJEMPLO 16.4

En la tabla 16.5 se muestran los valores mensuales del INPC de julio de 2009 a junio de 2011, mismos que se graficaron en la figura 16.1, donde se aprecia que tienen un comportamiento aproximadamente lineal, de manera que resulta razonable utilizar el modelo lineal de regresión para ajustar la tendencia.

Tabla 16.5 INPC mensual

Jul. 2009	93.671601856385
Ago. 2009	93.895719694096
Sep. 2009	94.366711949963
Oct. 2009	94.652203595540
Nov. 2009	95.143194058464
Dic. 2009	95.536951859488
Ene. 2010	96.575479439774
Feb. 2010	97.134050050685
Mar. 2010	97.823643397489
Abr. 2010	97.511947204733
Mayo 2010	96.897519532732
Jun. 2010	96.867177425472
Jul. 2010	97.077503396247
Ago. 2010	97.347134394847
Sep. 2010	97.857433471482
Oct. 2010	98.461517243282
Nov. 2010	99.250412032025
Dic. 2010	99.742092088296
Ene. 2011	100.228000000000
Feb. 2011	100.604000000000
Mar. 2011	100.797000000000
Abr. 2011	100.789000000000
Mayo 2011	100.046000000000
Jun. 2011	100.041000000000

Fuente: www.banxico.gob.mx

Solución: En la tabla 16.6 se muestran las operaciones necesarias para determinar la ecuación de regresión por el método de promedios y sumas de cuadrados que se revisó en la sección 13.3.3. Nótese que se numeraron los meses y se utilizaron estos enteros como la variable X.

Tabla 16.6 Datos y operaciones para el ejemplo 16.4

Mes	INPC	INPC		
X	Y	Y	$(X-\bar{X})^2$	$(X-\bar{X})(Y-\bar{n})$
Jul. 2009	1	93.6716	132.25	45.1369
Ago. 2009	2	93.89572	110.25	38.85872
Sep. 2009	3	94.36671	90.25	30.68346
Oct. 2009	4	94.6522	72.25	25.02694
Nov. 2009	5	95.14319	56.25	18.40017
Dic. 2009	6	95.53695	42.25	13.38739
Ene. 2010	7	96.57548	30.25	5.615888
Feb. 2010	8	97.13405	20.25	2.08125
Mar. 2010	9	97.82364	12.25	–0.79483
Abr. 2010	10	97.51195	6.25	0.211507
Mayo 2010	11	96.89752	2.25	1.048546
Jun. 2010	12	96.86718	0.25	0.364686
Jul. 2010	13	97.0775	0.25	–0.25952
Ago. 2010	14	97.34713	2.25	–0.37412
Sep. 2010	15	97.85743	6.25	0.652209
Oct. 2010	16	98.46152	12.25	3.027385
Nov. 2010	17	99.25041	20.25	7.442379
Dic. 2010	18	99.74209	30.25	11.80048
Ene. 2011	19	100.228	42.25	17.10443
Feb. 2011	20	100.604	56.25	22.55588
Mar. 2011	21	100.797	72.25	27.20383
Abr. 2011	22	100.789	90.25	30.32828
Mayo 2011	23	100.046	110.25	25.71923
Jun. 2011	24	100.041	132.25	28.11118
Promedios	12.5	97.59655		
		Sumas	1 150	353.3322

Entonces

$$SC_{XY} = 353.3322$$
$$SC_{XX} = 1\,150$$

De donde

$$\hat{\beta}_1 = \frac{SC_{XY}}{SC_{XX}} = \frac{353.3322}{1150} = 0.30725$$

Por lo que

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = 97.59655 - 0.30725(12.50) = 93.756.$$

La ecuación de regresión es

$$\hat{y} = 93.756 + 0.30725x$$

En la figura 16.4 se trazan los datos originales junto con la recta de regresión; como se vio en el capítulo 13 se puede utilizar esta ecuación para hacer pronósticos sobre el posible valor futuro de este INPC.

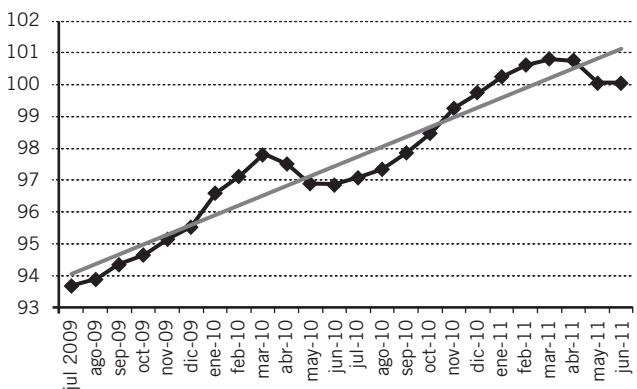


Figura 16.4 Nube de puntos y recta de regresión para el INPC.

Por supuesto, no todas las series de tiempo muestran un comportamiento rectilíneo, por lo que en ocasiones es necesario ajustar curvas que describan de mejor manera el comportamiento de la serie. Dos de las curvas que más comúnmente se presentan son la parábola y la curva exponencial. En las secciones siguientes se detallan los métodos para ajustar curvas a series de datos, mediante el método de mínimos cuadrados.

16.3.3 Ajuste de una función exponencial con mínimos cuadrados

Otras formas de curvas que son comunes en situaciones económicas y de negocios son las funciones exponenciales, que son de la forma:

$$y = f(x) = a^x \qquad \textbf{(16.5)}$$

en donde la base a es una constante positiva.

En la figura 16.5 se muestra una gráfica que ilustra la forma que tiene una función exponencial cuando $a > 1$, y en la figura 16.6 se muestra la gráfica de una función exponencial con $a < 1$.

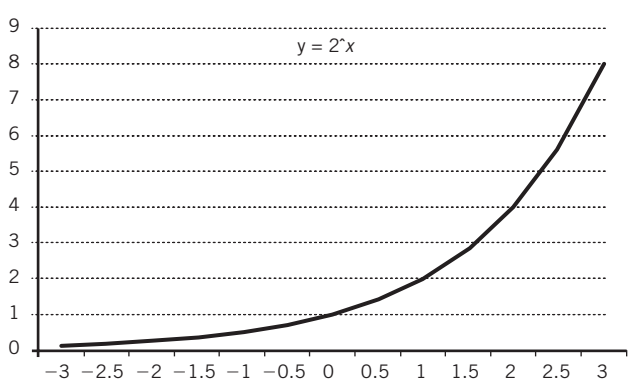


Figura 16.5 Ejemplo de función exponencial cuando $a > 1$.

El procedimiento que se utiliza para ajustar líneas de mínimos cuadrados a estas curvas consiste en ***linealizarlas***, es decir, convertir la ecuación exponencial en una lineal, a través del uso de logaritmos para, en un segundo paso, obtener la recta de mínimos cuadrados a partir de esta ecuación linealizada y, en un tercer paso, volver a su forma original (curva) a esa ecuación de regresión linealizada, convirtiéndola a su forma original a través de antilogaritmos. Se ilustra en el ejemplo siguiente el procedimiento.

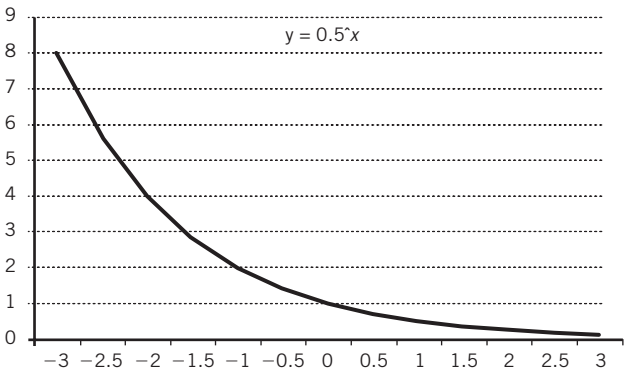


Figura 16.6 Ejemplo de función exponencial cuando $a < 1$.

Ejemplo 16.5

En las 2 primeras columnas de la tabla 16.7 se reproducen los datos de cierre de año del IPC de la BMV, de 1985 a 2010, ya vistos en el ejemplo 16.1, y en la figura 16.7 se muestra la gráfica correspondiente.

Linealizar. Convertir la ecuación exponencial en una lineal para obtener la recta de mínimos cuadrados y volver la ecuación de regresión linealizada a su forma original (curva).

Tabla 16.7 Cierres de año del IPC de la BMV

Año x	IPC y	ln y
1985	11.082	2.405322
1986	47.101	3.852294
1987	105.670	4.660321
1988	211.532	5.354376
1989	418.925	6.037692
1990	628.790	6.443797
1991	1 431.460	7.26645
1992	1 759.440	7.472751
1993	2 602.630	7.864278
1994	2 375.660	7.773031
1995	2 778.470	7.929656
1996	3 361.030	8.120003
1997	5 229.350	8.562042
1998	3 959.660	8.283913
1999	7 129.880	8.87205
2000	5 652.190	8.639798
2001	6 372.280	8.759713
2002	6 127.090	8.720475
2003	8 795.280	9.08197
2004	12 917.880	9.466368
2005	17 802.710	9.787106
2006	26 448.320	10.18295
2007	29 536.830	10.29339
2008	22 380.320	10.01594
2009	32 120.470	10.37725
2010	38 550.790	10.55973

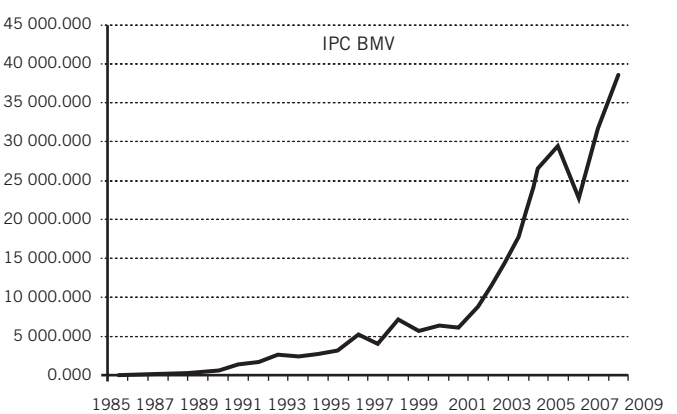


Figura 16.7 Gráfica de cierres de año del IPC de la BMV, 1985-2010.

Se puede apreciar que estos valores del IPC se asemejan a una función exponencial con $a > 1$.

Excel Para linealizar los datos se calculan los logaritmos naturales de los valores de la variable dependiente, y, con lo que se obtienen los datos de la tercera columna de la tabla 16.7. Para convertir los valores y en sus logaritmos naturales se utiliza la función "ln" de Excel. Si se colocan los datos originales

en una hoja de Excel a partir de la celda A1 y se introduce "=LN(B1)" en la celda C1 para, después, copiar esta función hacia abajo hasta abarcar todos los datos, se obtiene la información de la tercera columna.

Al aplicar ahora el procedimiento de "Regresión" de la sección de "Análisis de datos" de la pestaña de "Datos" de Excel, a esos datos de año y ln *y*, se obtienen los parámetros de la ecuación de regresión, como se muestra en la tabla 16.8, que es una versión recortada de la que Excel produce para ahorrar espacio y considerando que lo que interesa aquí son los parámetros de la ecuación lineal de mínimos cuadrados. Este procedimiento de regresión se explicó en detalle en la sección 13.3.4 del capítulo 13.

Tabla 16.8 Resultados de "Regresión" de Excel para los logaritmos naturales del ejemplo 16.5

Estadísticas de la regresión					
Coeficiente de correlación múltiple			0.93479078		
Coeficiente de determinación R^2			0.8738338		
R^2 ajustado			0.86857688		
Error típico			0.7578284		
Observaciones			26		
Análisis de varianza					
	Grados de libertad	Suma de cuadrados	Promedio de los cuadrados	*F*	Valor crítico de *F*
Regresión	1	95.4638226	95.4638226	166.225276	2.7895E–12
Residuos	24	13.7832933	0.57430389		
Total	25	109.247116			
	Coeficientes	Error típico	Estadístico *t*	Probabilidad	Inferior 95%
Intercepción	–502.385236	39.5833722	–12.6918251	3.883E–12	–584.0813
Año	0.25548857	0.01981632	12.8928382	2.7895E–12	0.2145897

Se puede ver en la esquina inferior izquierda de la tabla que el valor de la ordenada al origen es –502.385236, y que el valor de la pendiente es 0.25548857, por lo que la ecuación de regresión lineal es:

$$\hat{y} = -502.385236 + 0.25548857x$$

Ahora se calculan los valores $\hat{y}$ para todos los años de la serie, con lo que se obtienen los datos de la tercera columna de la tabla 16.9. En esta misma tabla se anotan en la columna 4 los antilogaritmos naturales de los valores de $\hat{y}$, con los que se les devuelve a sus unidades originales.

Tabla 16.9 Datos y cálculos para el ejemplo 16.5

Año *x*	IPC *y*	ln $\hat{y}$	$\hat{y}$
1985	11.082	4.759575	116.696
1986	47.101	5.015064	150.6653
1987	105.670	5.270553	194.5227
1988	211.532	5.526041	251.1468
1989	418.925	5.78153	324.2536
1990	628.790	6.037018	418.6412
1991	1 431.460	6.292507	540.50
1992	1 759.440	6.547995	697.84
1993	2 602.630	6.803484	900.98
1994	2 375.660	7.058973	1 163.24
1995	2 778.470	7.314461	1 501.85
1996	3 361.030	7.56995	1 939.03
1997	5 229.350	7.825438	2 503.47
1998	3 959.660	8.080927	3 232.21
1999	7 129.880	8.336415	4 173.08
2000	5 652.190	8.591904	5 387.83
2001	6 372.280	8.847393	6 956.19
2002	6 127.090	9.102881	8 981.08
2003	8 795.280	9.35837	11 595.40
2004	12 917.880	9.613858	14 970.72
2005	17 802.710	9.869347	19 328.58
2006	26 448.320	10.12484	24 954.98
2007	29 536.830	10.38032	32 219.17
2008	22 380.320	10.63581	41 597.92
2009	32 120.470	10.8913	53 706.74
2010	38 550.790	11.14679	69 340.35

Ahora, en la figura 16.8 se grafican tanto los datos originales del IPC, como los datos ajustados mediante mínimos cuadrados.

Como puede verse en esta figura, el ajuste parece bueno hasta 2007, cuando se da un brusco quiebre a la baja. Para mostrar un mejor ajuste, en el ejemplo siguiente se reducen los datos a los que gráficamente parecen ser una curva exponencial más suave, sin quiebres drásticos como la anterior.

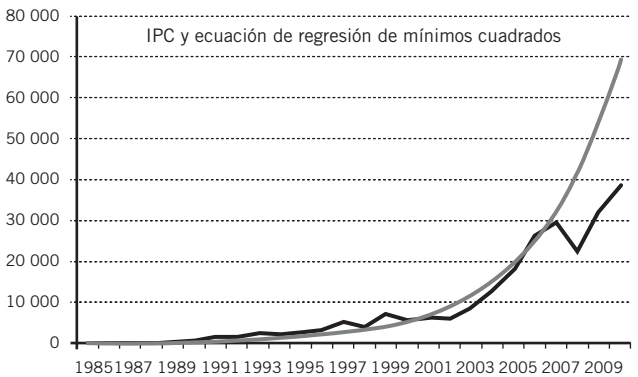


Figura 16.8 Gráficas del IPC y de su ecuación de regresión ajustada mediante mínimos cuadrados.

Ejemplo 16.6

Ajustar una función exponencial por el método de mínimos cuadrados a los datos del IPC, de 1991 a 2007.

Solución: En la tabla 16.10 se resumen las mismas operaciones que se llevaron a cabo en el ejemplo anterior con toda la serie.

Tabla 16.10 Datos y cálculos para el ejemplo 16.6

Año (1)	IPC y (2)	ln y (3)	ln $\hat{y}$ (4)	Anti ln $\hat{y}$ $\hat{y}$ (5)
1991	1 431.46	7.26645	7.272063	1 439.51
1992	1 759.44	7.472751	7.444496	1 710.414
1993	2 602.63	7.864278	7.616928	2 032.299
1994	2 375.66	7.773031	7.789361	2 414.761
1995	2 778.47	7.929656	7.961793	2 869.198
1996	3 361.03	8.120003	8.134226	3 409.157
1997	5 229.35	8.562042	8.306658	4 050.732
1998	3 959.66	8.283913	8.479091	4 813.045
1999	7 129.88	8.87205	8.651523	5 718.819
2000	5 652.19	8.639798	8.823956	6 795.053
2001	6 372.28	8.759713	8.996389	8 073.824
2002	6 127.09	8.720475	9.168821	95 93.249
2003	8 795.28	9.08197	9.341254	11 398.62
2004	12 917.88	9.466368	9.513686	13 543.74
2005	17 802.71	9.787106	9.686119	16 092.56
2006	26 448.32	10.18295	9.858551	19 121.04
2007	29 536.83	10.29339	10.03098	22 719.46

Igual que en el ejemplo anterior, el primer paso consistió en calcular los logaritmos naturales de los valores del IPC, los que aparecen en la columna (3).

Excel En el segundo paso se corrió el mecanismo de "Regresión" de Excel con los datos de años y logaritmos naturales de y, columna (1) y (3), con lo que se obtuvieron los resultados que se muestran en la tabla 16.11.

En tercer lugar, con los resultados de Excel se determinó que la ecuación de regresión de mínimos cuadrados es:

$$\hat{y} = -336.041144 + 0.17243255x$$

En cuarto lugar, con esta ecuación de regresión se calcularon los valores estimados de $\hat{y}$, que son logaritmos y son los que aparecen en la columna (4) de la tabla.

Tabla 16.11 El cuadro de resultados de "Regresión" de Excel para los datos de años y ln *y*

Estadísticas de la regresión					
Coeficiente de correlación múltiple			0.97006837		
Coeficiente de determinación R^2			0.94103264		
R^2 ajustado			0.93710148		
Error típico			0.22511643		
Observaciones			17		
Análisis de la varianza					
	Grados de libertad	**Suma de cuadrados**	**Promedio de los cuadrados**	***F***	**Valor crítico de *F***
Regresión	1	12.1310578	12.1310578	239.378019	1.2527E–10
Residuos	15	0.76016113	0.05067741		
Total	16	12.8912189			
	Coeficientes	**Error típico**	**Estadístico *t***	**Probabilidad**	**Inferior 95%**
Intercepción	–336.041144	22.2787704	–15.083469	1.795E–10	–383.527219
Año	0.17243255	0.01114492	15.471846	1.2527E–10	0.14867771

En un quinto paso se calcularon los antilogaritmos naturales de los valores ln $\hat{y}$, para encontrar los valores estimados del IPC en sus dimensiones originales y que son los de la columna (5) de la tabla.

Finalmente, en la figura 16.9 se grafican tanto los datos originales del IPC como los ajustados por mínimos cuadrados.

Aunque el ajuste parece mejor que el realizado en el ejemplo anterior con la serie completa de datos, aún restaría evaluar qué tan bueno es el ajuste. Sin embargo, esta tarea rebasa los alcances de este texto.

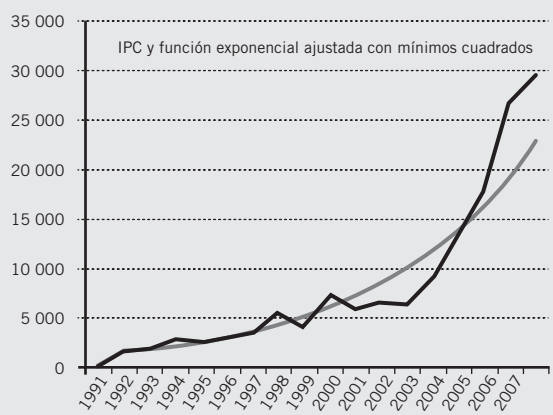


Figura 16.9 Gráfica del IPC y de la función exponencial ajustada mediante mínimos cuadrados.

16.3.4 Ajuste de una parábola con mínimos cuadrados

En esta sección se revisa la forma en la que una serie que presenta un comportamiento aproximadamente parabólico se ajusta con el método de los mínimos cuadrados y se comienza por revisar las que son ahora 3 ecuaciones normales con las que se construye la ecuación de regresión cuadrática.

Se comienza por derivar las ecuaciones normales cuya resolución simultánea permite calcular los parámetros de una ecuación de segundo grado, la que permite ajustar una parábola:

Si $y_1, y_2, \ldots y_n$ representan a las observaciones de la variable y, $x_1, x_2, \ldots x_n$ representan las observaciones de la variable x, y si

$$y_1 = a + bx_1 + cx_1^2$$
$$y_2 = a + bx_2 + cx_2^2$$
$$\ldots$$

la suma de estas expresiones es:

$$y_1 = a + bx_1 + cx_1^2$$
$$+$$
$$y_2 = a + bx_2 + cx_2^2$$
$$+$$
$$\cdots$$
$$y_n = a + bx_n + cx_n^2$$

Es igual a:

$$\sum y = na + b\sum x + c\sum x^2$$

Si se multiplica cada una de las ecuaciones de la forma $y = a + bx + cx^2$ por el coeficiente de la primera incógnita de la ecuación (1, que es el coeficiente de la primera incógnita, a), no las altera, por lo que la suma de esas ecuaciones resultantes (que no cambiaron) es igual a la suma anotada antes,

$$\sum y = na + b\sum x + c\sum x^2 \tag{16.6}$$

la que se conoce como "ecuación normal I".

Si ahora se multiplica cada una de las ecuaciones de la forma $y = a + bx + cx^2$ por el coeficiente de la segunda incógnita de la ecuación (x, que es el coeficiente de la segunda incógnita, b), se obtiene:

$$x_1y_1 = ax_1 + bx_1^2 + cx_1^3$$
$$+$$
$$x_2y_2 = ax_2 + bx_2^2 + cx_2^3$$
$$+$$
$$\cdots$$
$$x_ny_n = ax_n + bx_n^2 + cx_n^3$$

Sumando ahora estas ecuaciones se llega a la ecuación normal II:

$$\sum x_iy_i = a\sum x_i + b\sum x_i^2 + c\sum x_i^3 \tag{16.7}$$

Ahora, para obtener la ecuación normal III, en primer lugar se multiplica cada una de las ecuaciones de la forma $y = a + bx + cx^2$ por el coeficiente de la tercera incógnita, x^2, se tiene:

$$x_1^2y_1 = ax_1^2 + bx_1^3 + cx_1^4$$
$$+$$
$$x_2^2y_2 = ax_2^2 + bx_2^3 + cx_2^4$$
$$+$$
$$\cdots$$
$$x_n^2y_n = ax_n^2 + bx_n^3 + cx_n^4$$

las cuales, sumadas, producen:

$$\sum x_i^2y_i = a\sum x_i^2 + b\sum x_i^3 + c\sum x_i^4 \tag{16.8}$$

que es lo que se conoce como la ecuación normal III.

Resumiendo el conjunto de las ecuaciones normales cuya resolución simultánea permite construir la ecuación de mínimos cuadrados de una parábola

$$\sum y = na + b\sum x + c\sum x^2 \tag{16.6}$$
$$\sum x_iy_i = a\sum x_i + b\sum x_i^2 + c\sum x_i^3 \tag{16.7}$$
$$\sum x_i^2y_i = a\sum x_i^2 + b\sum x_i^3 + c\sum x_i^4 \tag{16.8}$$

Se muestra ahora en el ejemplo siguiente la forma en la que se utilizan estas ecuaciones normales para ajustar una parábola a una nube de puntos.

Ejemplo 16.7

En la tabla 16.12 se muestran los valores del tipo de cambio del dólar respecto al peso mexicano entre diciembre de 2008 y abril de 2009 tomando en cuenta su valor cada 5 días.

Tabla 16.12 Tipo de cambio peso-dólar en intervalos de 5 días

05 dic. 2008	13.5358
10 dic. 2008	13.4921
15 dic. 2008	13.1821
20 dic. 2008	13.2150
25 dic. 2008	13.2175
30 dic. 2008	13.5383
04 ene. 2009	13.8325
09 ene. 2009	13.5045
14 ene. 2009	13.8390
19 ene. 2009	13.9058
24 ene. 2009	13.9325
29 ene. 2009	14.1513
03 feb. 2009	14.3097
08 feb. 2009	14.2945
13 feb. 2009	14.6013
18 feb. 2009	14.6118
23 feb. 2009	14.8163
28 feb. 2009	14.9322
05 mar. 2009	15.2178
10 mar. 2009	15.3517
15 mar. 2009	14.9083
20 mar. 2009	14.0800
25 mar. 2009	14.2683
30 mar. 2009	14.3317
04 abr. 2009	13.7924
09 abr. 2009	13.5712
14 abr. 2009	13.0914
19 abr. 2009	13.0511

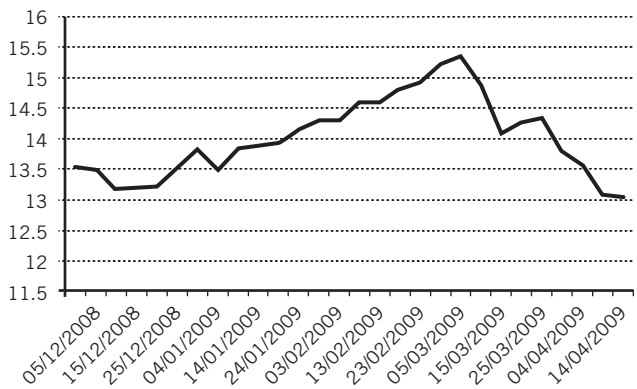


Figura 16.10 Gráfica de los datos del ejemplo 16.7.

Solución: En la figura 16.10 se muestra la gráfica correspondiente, en donde puede apreciarse que su comportamiento es aproximadamente el de una parábola cuya forma algebraica general es de la forma $y = a + bx + cx^2$.

En la tabla 16.13 se muestra el procedimiento correspondiente para determinar la ecuación por el método de promedios y sumas de cuadrados que también fue utilizado en el ejemplo 16.4. Para este ejercicio se numeraron las fechas y se utiliza como la variable X.

Tabla 16.13 Datos y operaciones para el ejemplo 16.2

	Mes					Tipo de cambio		
	X	Y	x^2	x^3	x^4	y	xy	x^2y
05 dic. 2008	1	13.5358	1	1	1	13.5358	13.54	13.536
10 dic. 2008	2	13.4921	4	8	16	13.4921	26.98	53.968
15 dic. 2008	3	13.1821	9	27	81	13.1821	39.55	118.639
20 dic. 2008	4	13.2150	16	64	256	13.215	52.86	211.440
25 dic. 2008	5	13.2175	25	125	625	13.2175	66.09	330.438
30 dic. 2008	6	13.5383	36	216	1 296	13.5383	81.23	487.379
04 ene. 2009	7	13.8325	49	343	2 401	13.8325	96.83	677.793
09 ene. 2009	8	13.5045	64	512	4 096	13.5045	108.04	864.288
14 ene. 2009	9	13.8390	81	729	6 561	13.839	124.55	1 120.959
19 ene. 2009	10	13.9058	100	1 000	10 000	13.9058	139.06	1 390.580
24 ene. 2009	11	13.9325	121	1 331	14 641	13.9325	153.26	1 685.833
29 ene. 2009	12	14.1513	144	1 728	20 736	14.1513	169.82	2 037.787
03 feb. 2009	13	14.3097	169	2 197	28 561	14.3097	186.03	2 418.339
08 feb. 2009	14	14.2945	196	2 744	38 416	14.2945	200.12	2 801.722
13 feb. 2009	15	14.6013	225	3 375	50 625	14.6013	219.02	3 285.293

	Mes					Tipo de cambio		
	x	Y	x^2	x^3	x^4	y	xy	x^2y
18 feb. 2009	16	14.6118	256	4 096	65 536	14.6118	233.79	3 740.621
23 feb. 2009	17	14.8163	289	4 913	83 521	14.8163	251.88	4 281.911
28 feb. 2009	18	14.9322	324	5 832	104 976	14.9322	268.78	4 838.033
05 mar. 2009	19	15.2178	361	6 859	130 321	15.2178	289.14	5 493.626
10 mar. 2009	20	15.3517	400	8 000	160 000	15.3517	307.03	6 140.680
15 mar. 2009	21	14.9083	441	9 261	194 481	14.9083	313.07	6 574.560
20 mar. 2009	22	14.0800	484	10 648	234 256	14.08	309.76	6 814.720
25 mar. 2009	23	14.2683	529	12 167	279 841	14.2683	328.17	7 547.931
30 mar. 2009	24	14.3317	576	13 824	331 776	14.3317	343.96	8 255.059
04 abr. 2009	25	13.7924	625	15 625	390 625	13.7924	344.81	8 620.250
09 abr. 2009	26	13.5712	676	17 576	456 976	13.5712	352.85	9 174.131
14 abr. 2009	27	13.0914	729	19 683	531 441	13.0914	353.47	9 543.631
19 abr. 2009	28	13.0511	784	21 952	614 656	13.0511	365.43	10 232.062
Sumas	406		7 714	164 836	3 756 718	393	5 739	108 755

Entonces,

$$393 = 28a + 406b + 7\,714c$$
$$5\,739 = 406a + 7\,714b + 164\,836c$$
$$108\,755 = 7\,714a + 164\,836b + 3\,756\,718c.$$

De donde

$$a = 12.6596$$
$$b = 0.2331$$
$$c = -0.0073$$
$$y = a + bx + cx^2 = 12.6596 + 0.2331x - 0.0073x^2$$

En la figura 16.11 se graficaron los datos originales con una parábola.

Aquí, de nuevo, sería necesario evaluar qué tan bueno es el ajuste, ya que en principio, y a partir de la observación de la gráfica, el ajuste no parece ser muy estrecho. Sin embargo, las técnicas para hacer esto rebasan el alcance de este texto.

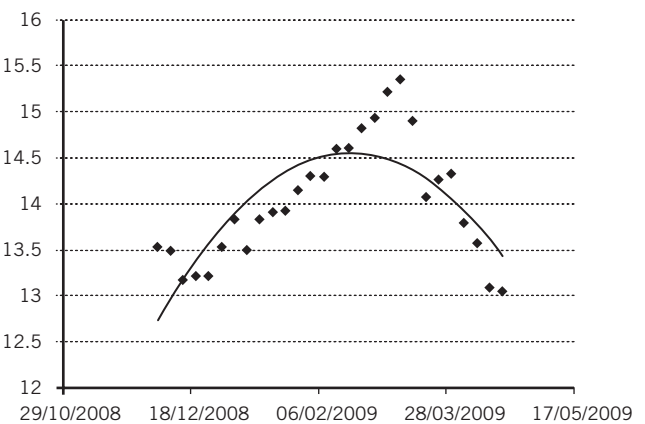


Figura 16.11 Nube de puntos y parábola para el tipo de cambio.

ejercicios 16.3 Tendencia secular

Ajuste y pronósticos con promedios móviles

1. Ingrese a la sección de finanzas de Yahoo! (http://mx.finance.yahoo.com) en internet y obtenga datos históricos para los precios de alguna de las acciones que cotizan en la Bolsa Mexicana de Valores (BMV). Algunos ejemplos de las claves que se pueden anotar en el globo de "Buscar cotizaciones" de dicha página web para encontrar cotizaciones de acciones mexicanas que cotizan en esa BMV son (a todos ellos se les debe añadir ".mx" al final):

ALAA	CHDRAUIB	GMOELOC	OHLEMEX
AMXL	COMERCIUBC	GRUAB	TELEXL

ARAL*	COMPARC	HMEX	TLEVISCPO
ASURB	ELEKTRA	ICB*	URB
AUTLAB	FEMSAUBD	KIMBERA	WALMEXV
AZTEACPO	GAPB	KOFL	
BIMOA	GEOB	LABB	
BOLSAA	GFINBURO	MEXCEM	
C	GFNORTEO	MFRISCOA1	
CEMECPO	GMEXCOB	NAFTRAC02	

a) Ingrese a la sección de "Precios históricos" cuya liga está del lado izquierdo de la página de cada acción y obtenga cuando menos 100 de las observaciones más recientes.

b) Obtenga promedios móviles de los precios de cierre con periodo de 5 y de 20 días y grafíquelos junto con los datos de los precios de cierre.

c) Observe la gráfica y describa cómo se comportan los precios de cierre cuando los 2 promedios móviles se cruzan.

2. En la tabla siguiente se presentan datos de personas aseguradas en el Instituto Mexicano del Seguro Social (IMSS).

Año	Asegurados en el IMSS (miles de personas)
1995	38 575
1996	38 953
1997	37 465
1998	36 738
1999	36 554
2000	34 324
2001	37 261
2002	39 462
2003	41 942
2004	44 557
2005	46 534
2006	45 872
2007	46 199
2008	41 243
2009	44 961
2010	47 918
2011	50 561

Fuente: INEGI, "Anuario estadístico de los Estados Unidos Mexicanos", disponible en: www.inegi.org.mx, consultado el 24 de febrero de 2012.

a) Ensaye ajustes con una serie de promedios móviles exponenciales con diferentes factores de suavización hasta encontrar una que resulte satisfactoria y explicar cómo se llegó a ella.

b) Pronostique cuántos asegurados se esperaría que tenga el IMSS en 2011.

3. En la tabla siguiente se muestran datos de la población penitenciaria total.

Año	Población penitenciaria total
1995	91 422
1996	97 565
1997	99 858
1998	103 916
1999	108 808
2000	121 135
2001	134 567
2002	140 415
2003	147 809
2004	154 825
2005	159 628
2006	164 929
2007	165 970
2008	171 437
2009	173 060

Fuente: INEGI, "Anuario estadístico de los Estados Unidos Mexicanos", disponible en: www.inegi.org.mx, consultado el 24 de febrero de 2012.

a) Ensaye ajustes con una serie de promedios móviles exponenciales con diferentes factores de suavización hasta encontrar una que resulte satisfactorio y explicar cómo se llegó a ella.

b) Pronostique cuántos reos se esperaría que haya en 2015.

Ajuste de rectas y curvas con el método de mínimos cuadrados

4. En análisis bursátil se utilizan mediciones de la variabilidad de los precios (la desviación estándar, por lo general) para ajustar los periodos de cálculo de diversos promedios móviles, de manera que se calculan dichos promedios con periodos variables según esa volatilidad de los precios. En la tabla que aparece en seguida se muestran los resultados para esos valores de los periodos móviles aplicados al IPC de determinado periodo.

Núm. de días en el promedio (longitud efectiva)
49.00
24.00
15.67
11.50
9.00
7.33
6.14
5.25
4.56
4

Ajuste por el método de mínimos cuadrados la curva o la recta más apropiada

5. Ingrese al sitio del Banco de México en internet (www.banxico.gob.mx) y continúe con las siguientes ligas:

- Estadísticas.
- Política monetaria e inflación.

- Financiamiento e información financiera de intermediarios financieros.
- En la sección de Sistema Bancario.
- Crédito por actividad principal de prestatarios.
- Banca comercial.
- Banca comercial crédito por actividad principal de prestatarios.

Dé clic sobre el ícono "XLS" que aparece en la parte superior izquierda de la página; entonces se abrirá un archivo de Excel con 8 columnas que contiene datos sobre, precisamente, crédito por actividad principal de prestatarios de la banca comercial. La primera columna corresponde al periodo y las columnas restantes tienen los siguientes encabezados:

- CF29 Banca Comercial Total (I a X).
- CF29 Banca Comercial Total (I a X), Sector privado del país (I a VI) - Empresas y personas físicas.
- CF29 Banca Comercial Total (I a X), VII. Sector financiero del país (A+B) 4/.
- CF29 Banca Comercial Total (I a X), VIII. Sector público.
- CF245 Banca comercial cartera vigente (I a X), IX. OTROS (A+B+C).
- CF29 Banca Comercial Total (I a X), X. Sector externo (A+B).
- CF29 Banca Comercial Total (I a X), XI. Crédito intrabancario (A+B) 4/.

En la parte final de cada uno de estos encabezados se puede distinguir a qué se refieren los datos correspondientes.

a) Con los datos de la Banca Comercial Total, ajuste por el método de mínimos cuadrados la curva o la recta más apropiada.
b) Con los datos del crédito intrabancario, ajuste por el método de mínimos cuadrados la curva o la recta más apropiada.
c) Con los datos de las otras columnas, ajuste por el método de mínimos cuadrados la curva o la recta más apropiada.
d) Comentar sobre el comportamiento de las diferentes series y sus implicaciones para la economía nacional.

6. Ingrese al sitio del Banco de México en internet (www.banxico.gob.mx) y continúe con las siguientes ligas:

- Estadísticas.
- Sistema financiero.
- Mercados financieros (tipo de cambio, tasas de interés y derivados).

En la sección "Mercado cambiario":

- Tipos de cambio.
- Cotización de las divisas que conforman la canasta de DEG[2]…

Dé clic sobre el ícono "XLS" que aparece en la parte superior izquierda de la página; entonces se abrirá un archivo de Excel con varias columnas que contiene datos sobre, precisamente, cotizaciones respecto al peso mexicano del dólar estadounidense, el yen japonés, la libra esterlina, el euro y los propios DEG.

a) Con los datos completos de la cotización del peso respecto al euro, o alguna parte apropiada o de interés, ajuste por el método de mínimos cuadrados la curva o la recta más apropiada.
b) Con los datos completos de la cotización del peso respecto al dólar, o alguna parte apropiada o de interés, ajuste por el método de mínimos cuadrados la curva o la recta más apropiada.
c) Con los datos de las otras columnas, ajuste por el método de mínimos cuadrados la curva o la recta más apropiada.
d) Comente sobre el comportamiento de las diferentes series y sus implicaciones.

7. Ajuste a los siguientes datos sobre tasa bruta de natalidad por 1 000 habitantes, por el método de mínimos cuadrados la curva o la recta más apropiada y, con base en su ecuación, estime cuál será la tasa de natalidad para 2015.

Año	Tasa bruta de natalidad (por 1 000 habitantes)
1995	25.8
1996	25.0
1997	24.2
1998	24.0
1999	24.2
2000	23.7
2001	22.9
2002	21.6
2003	20.5
2004	19.7
2005	19.3
2006	19.0
2007	18.6
2008	18.3
2009	18.0
2010	17.8
2011	17.5

Fuente: INEGI, "Anuario estadístico de los Estados Unidos Mexicanos", disponible en: www.inegi.org.mx, consultado el 24 de febrero de 2012.

[2] Las siglas DEG significan derechos especiales de giro; se sugiere buscar en internet qué son.

8. Ajuste a los siguientes datos sobre la tasa bruta de mortalidad por 1 000 habitantes, por el método de mínimos cuadrados la curva o la recta más apropiada y, con base en su ecuación, estime cuál será la tasa de mortalidad para 2015.

Año	Tasa bruta de mortalidad (por 1 000 habitantes)
1995	5.0
1996	4.9
1997	4.8
1998	4.7
1999	4.7
2000	4.7
2001	4.6
2002	4.6
2003	4.7
2004	4.7
2005	4.8
2006	4.8
2007	4.8
2008	4.9
2009	4.9
2010	5.0
2011	5.0

Fuente: INEGI, "Anuario estadístico de los Estados Unidos Mexicanos", disponible en: www.inegi.org.mx, consultado el 17 de octubre de 2011.

16.4 Variaciones estacionales

Este tipo de variaciones se refieren a variaciones de las estaciones del año, las cuales se relacionan con el clima: producen ciclos como los agrícolas o los patrones de compras que se observan en los consumidores según la temporada; prevén aumento de las ventas de ropa cálida en invierno o la abundancia o escasez de determinados productos agrícolas; sin embargo, la estacionalidad puede tener también otra periodicidad, como la semanal en el caso de un restaurante que ve subir sus ventas los fines de semana.

En las secciones siguientes se detalla uno de los procedimientos que se utilizan para obtener *índices estacionales*, los cuales pueden utilizarse para *a*) "desestacionalizar" series de tiempo y *b*) para hacer pronósticos sobre el comportamiento futuro de la serie.

16.4 1 Cálculo de índices estacionales

Se ilustra el procedimiento con el siguiente ejemplo.

EJEMPLO 16.8

En la tabla 16.14 se muestran los datos mensuales de ventas de gruesas de naranja (una gruesa equivale a 144 piezas) que realiza una cooperativa veracruzana que se dedica al cultivo del cítrico.

Tabla 16.14 Ventas mensuales de gruesas de naranja de una cooperativa veracruzana

	2007	2008	2009	2010	2011
Ene.	1 920	1 280	2 555	1 708	1 299
Feb.	2 080	1 296	2 480	1 596	1 413
Mar.	2 488	1 680	2 796	1 771	1 704
Abr.	2 328	1 440	1 700	1 816	1 332
Mayo	1 624	820	1 084	1 245	1 148
Jun.	514	213	359	715	340
Jul.	232	115	117	471	95
Ago.	56	44	26	147	12
Sep.	392	646	670	264	184
Oct.	1 920	2 772	1 892	1 214	869
Nov.	1 680	2 214	1 432	1 220	1 194
Dic.	1 440	2 569	1 995	1 840	1 314

Solución: Una primera forma de apreciar la estacionalidad de estos datos consiste en graficarlos año por año, como en la figura 16.12.

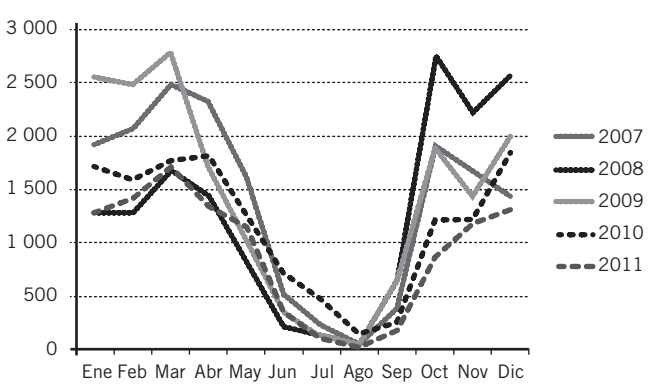


Figura 16.12 Gráficas de las ventas anuales de naranja.

En esta gráfica se puede apreciar cómo las ventas se reducen marcadamente entre julio y septiembre de cada año.

Con el propósito de eliminar las variaciones estacionales se construyen índices estacionales mensuales para los datos de gruesas de naranjas. El procedimiento es un tanto laborioso por lo que lo más conveniente es utilizar Excel para realizar las operaciones. Los pasos a seguir son:

1. Colocar los datos en una hoja de Excel, todos en la misma columna, con un renglón en blanco entre cada par de meses, como se muestra en las 3 primeras columnas de la tabla 16.15, la cual también contiene otras operaciones, mismas que se describen en los pasos siguientes.

Tabla 16.15 Datos y operaciones para el ejemplo 16.8

Año (1)	Mes (2)	Ventas (gruesas) (3)	Totales móviles de 12 meses centrado a medio año (4)	Suma de 2 totales móviles consecutivos (5)	Promedio móvil con la suma de 2 totales móviles consecutivos (6)	Cociente entre datos (gruesas) y promedio móvil (7)
	Ene.	1 920				
	Feb.	2 080				
	Mar.	2 488				
	Abr.	2 328				
	Mayo	1 624				
	Jun.	514				
			16 674			
2007	Jul.	232		32 708	1 362.833	0.170234
			16 034			
	Ago.	56		31 284	1303.5	0.042961
			15 250			
	Sep.	392		29 692	1 237.167	0.316853
			14 442			
	Oct.	1 920		27 996	1 166.5	1.645949
			13 554			
	Nov.	1 680		26 304	1 096	1.532847
			12 750			
	Dic.	1 440		25 199	1 049.958	1.371483
			12 449			
	Ene.	1 280		24 781	1 032.542	1.239659
			12 332			
	Feb.	1 296		24 652	1 027.167	1.261723
			12 320			
	Mar.	1 680		24 894	1 037.25	1.619667
			12 574			
	Abr.	1 440		26 000	1 083.333	1.329231
2008			13 426			
	Mayo	820		27 386	1 141.083	0.718615
			13 960			
	Jun.	213		29 049	1 210.375	0.175979
			15 089			
	Jul.	115		31 453	1 310.542	0.08775
			16 364			
	Ago.	44		33 912	1 413	0.031139

(*continúa*)

Tabla 16.15 (*continuación*)

Año (1)	Mes (2)	Ventas (gruesas) (3)	Totales móviles de 12 meses centrado a medio año (4)	Suma de 2 totales móviles consecutivos (5)	Promedio móvil con la suma de 2 totales móviles consecutivos (6)	Cociente entre datos (gruesas) y promedio móvil (7)
			17 548			
	Sep.	646		36 212	1 508.833	0.428145
			18 664			
	Oct.	2 772		37 588	1 566.167	1.769927
2008			18 924			
	Nov.	2 214		38 112	1 588	1.394207
			19 188			
	Dic.	2 569		38 522	1 605.083	1.60054
			19 334			
	Ene.	2 555		38 670	1 611.25	1.585725
			19 336			
	Feb.	2 480		38 654	1 610.583	1.539815
			19 318			
	Mar.	2 796		38 660	1 610.833	1.735748
			19 342			
	Abr.	1 700		3 804	1 575.167	1.079251
			18 462			
	Mayo	1 084		36 142	1 505.917	0.719827
			17 680			
	Jun.	359		34 786	1 449.417	0.247686
2009			17 106			
	Jul.	117		33 365	1 390.208	0.08416
			16 259			
	Ago.	26		31 634	1 318.083	0.019726
			15 375			
	Sep.	670		29 725	1 238.542	0.540959
			14 350			
	Oct.	1 892		28 816	1 200.667	1.575791
			14 466			
	Nov.	1 432		29 093	1 212.208	1.181315
			14 627			
	Dic.	1 995		29 610	1 233.75	1.617021
			14 983			
	Ene.	1 708		30 320	1 263.333	1.351979
			15 337			
	Feb.	1 596		30 795	1 283.125	1.243838
			15 458			
	Mar.	1 771		30 510	1 271.25	1.393117
2010			15 052			
	Abr.	1 816		29 426	1 226.083	1.481139
			14 374			
	Mayo	1 245		28 536	1 189	1.047098
			14 162			
	Jun.	715		28 169	1 173.708	0.60918
			14 007			

Año (1)	Mes (2)	Ventas (gruesas) (3)	Totales móviles de 12 meses centrado a medio año (4)	Suma de 2 totales móviles consecutivos (5)	Promedio móvil con la suma de 2 totales móviles consecutivos (6)	Cociente entre datos (gruesas) y promedio móvil (7)
2010	Jul.	471		27 605	1 150.208	0.409491
			13 598			
	Ago.	147		27 013	1 125.542	0.130604
			13 415			
	Sep.	264		26 763	1 115.125	0.236745
			13 348			
	Oct.	1 214		26 212	1 092.167	1.111552
			12 864			
	Nov.	1 220		25 631	1 067.958	1.142367
			12 767			
	Dic.	1 840		25 159	1 048.292	1.755237
			12 392			
2011	Ene.	1 299		24 408	1 017	1.277286
			12 016			
	Feb.	1 413		23 897	995.7083	1.41909
			11 881			
	Mar.	1 704		23 682	986.75	1.726881
			11 801			
	Abr.	1 332		23 257	969.0417	1.374554
			11 456			
	Mayo	1 148		22 886	953.5833	1.20388
			11 430			
	Jun.	340		22 334	930.5833	0.365362
			10 904			
	Jul.	95				
	Ago.	12				
	Sep.	184				
	Oct.	869				
	Nov.	1 194				
	Dic.	1 314				

2. En la columna (4) se calculan totales de 12 meses centrados a medio año. Por ejemplo, el primer valor de esta columna (16 674) que se anotó entre los meses de junio y julio de 2007, es la suma de las ventas de los 12 meses de ese año, el total que sigue incluye los 11 últimos meses de 2007 y el primero de 2008, y así sucesivamente.
3. En la columna (5) se anotan las sumas de 2 totales móviles consecutivos de la columna (4). Nótese, para empezar que, como los totales de la columna (4) están en los renglones intermedios entre meses, las sumas de esta columna (5) caen directamente en los meses. Así, el primer valor de esta columna, 32 708, es la suma de las 2 cantidades de la columna (4) entre las que se encuentra: 16 674 y 16 034. Las demás sumas se calculan de la misma manera.
4. En la columna (6) se anota el resultado de dividir la suma de la columna (5) entre 24. Se le llama "Promedio móvil con la suma de 2 totales móviles consecutivos" porque, considerando que el valor de la columna (5) es la suma de 2 totales anuales, entonces incluye 24 meses.
5. Para obtener los valores de la columna (7) se dividen los valores originales de las ventas en gruesas de naranjas entre los promedios móviles de la columna (6). Comenzando

con el primer valor, 0.170234, se obtuvo dividiendo las 232 gruesas de naranjas vendidas en julio de 2007, columna (3) entre 1 362.833, el correspondiente promedio móvil de la columna 6.

6. El siguiente paso consiste en reacomodar los cocientes de la columna (7), de manera que los valores de cada mes queden en el mismo renglón para todos los años, como se muestra en la tabla 16.16.

Tabla 16.16 Resumen de cocientes de la columna (7) de la tabla de datos para el ejemplo 16.8

Año	Mes	Cociente entre datos (gruesas) y promedio móvil				
		2007	2008	2009	2010	2011
2007	Ene.		1.239659	1.585725	1.351979	1.277286
	Feb.		1.261723	1.539815	1.243838	1.41909
	Mar.		1.619667	1.735748	1.393117	1.726881
	Abr.		1.329231	1.079251	1.481139	1.374554
	Mayo		0.718615	0.719827	1.047098	1.20388
	Jun.		0.175979	0.247686	0.60918	0.365362
	Jul.	0.170234	0.08775	0.08416	0.409491	
	Ago.	0.042961	0.031139	0.019726	0.130604	
	Sep.	0.316853	0.428145	0.540959	0.236745	
	Oct.	1.645949	1.769927	1.575791	1.111552	
	Nov.	1.532847	1.394207	1.181315	1.142367	
	Dic.	1.371483	1.60054	1.617021	1.755237	

7. Ahora se deben eliminar los valores máximo y mínimo de cada renglón. Si se asume que la tabla está colocada en el extremo izquierdo superior de la hoja de Excel, se pueden anotar en Excel las funciones "=MAX(C3:G3)", y "=MIN(C3:G3)", en las columnas H e I, respectivamente, y copiarlas hacia abajo para obtener, precisamente esos valores máximo y mínimo por renglón.

Figura 16.13 Cuadro de diálogo de "Pegado especial" de Excel.

Como se deben eliminar estos valores en cada renglón, se deben copiar todos los máximos y mínimos obtenidos, para volverlos a copiar en el mismo lugar, pero con una operación de "Pegado especial", en el que se elija la opción de "Valores" del cuadro de diálogo de Excel que aparece cuando se inicia la operación de copiado especial. Se muestra este diálogo en la figura 16.13.

Con esta operación de pegado especial, Excel convierte lo que originalmente eran las fórmulas que se introdujeron para calcular los máximos y los mínimos en los valores mismos y, con esto ya se pueden eliminar en cada renglón estos valores. Si no se hace esto, al eliminarlos, se altera la fórmula y, con ello, toda la tabla.

En la tabla 16.17 se realizaron estas operaciones, de manera que en las 2 últimas columnas se tienen los valores máximo y mínimo de cada renglón.

El lector ya se habrá percatado de que, como las tablas anteriores tenían renglones en blanco, fue necesario eliminar datos innecesarios que se crean al "correr" o "deslizar" las fórmulas de Excel para repetir cálculos. En lo siguiente, como ya no se requieren huecos entre los meses, conviene eliminar esos renglones en blanco (tablas 16.10 y siguientes).

Tabla 16.17 Máximos y mínimos por renglón para el ejemplo 16.8

Año	Mes	Cociente entre datos (gruesas) y promedio móvil					Máximo	Mínimo
		2007	2008	2009	2010	2011		
2007	Ene.		1.239659	1.585725	1.351979	1.277286	1.585725	1.239659
	Feb.		1.261723	1.539815	1.243838	1.41909	1.539815	1.243838
	Mar.		1.619667	1.735748	1.393117	1.726881	1.735748	1.393117

Año	Mes	Cociente entre datos (gruesas) y promedio móvil					Máximo	Mínimo
		2007	2008	2009	2010	2011		
2007	Abr.		1.329231	1.079251	1.481139	1.374554	1.481139	1.079251
	Mayo		0.718615	0.719827	1.047098	1.20388	1.20388	0.718615
	Jun		0.175979	0.247686	0.60918	0.365362	0.60918	0.175979
	Jul.	0.170234	0.08775	0.08416	0.409491		0.409491	0.08416
	Ago.	0.042961	0.031139	0.019726	0.130604		0.130604	0.019726
	Sep.	0.316853	0.428145	0.540959	0.236745		0.540959	0.236745
	Oct.	1.645949	1.769927	1.575791	1.111552		1.769927	1.111552
	Nov.	1.532847	1.394207	1.181315	1.142367		1.532847	1.142367
	Dic.	1.371483	1.60054	1.617021	1.755237		1.755237	1.371483

8. Con la eliminación de los valores máximo y mínimo por renglón se elimina la influencia de los valores extremos. En la tabla 16.18 se muestran los resultados de esta operación, ya sin los datos de máximos y mínimos.

Tabla 16.18 Datos de cocientes para el ejemplo 16.8, eliminados los máximos y los mínimos de renglón

Año	Mes	Cociente entre datos (gruesas) y promedio móvil				
		2007	2008	2009	2010	2011
2007	Ene.				1.351979	1.277286
	Feb.		1.261723			1.41909
	Mar.		1.619667			1.726881
	Abr.		1.329231			1.374554
	Mayo			0.719827	1.047098	
	Jun.			0.247686		0.365362
	Jul.	0.170234	0.08775			
	Ago.	0.042961	0.031139			
	Sep.	0.316853	0.428145			
	Oct.	1.645949		1.575791		
	Nov.		1.394207	1.181315		
	Dic.		1.60054	1.617021		

9. Una vez eliminados los valores máximo y mínimo de cada renglón se obtienen promedios por renglón, para lo cual basta con anotar la función "=PROMEDIO(C3:G3)" en cada renglón. Éstos son los datos que se muestran en la penúltima columna de la tabla 16.19.

Tabla 16.19 Cálculo de promedios de cocientes, eliminados los máximos y los mínimos por renglón

Año	Mes	Cociente entre datos (gruesas) y promedio móvil					Promedio	Promedio ajustado a 12 (índices estacionales)
		2007	2008	2009	2010	2011		
2007	Ene.				1.351979	1.277286	1.314633	1.323884
	Feb.		1.261723			1.41909	1.340407	1.349839
	Mar.		1.619667			1.726881	1.673274	1.685049
	Abr.		1.329231			1.374554	1.351892	1.361406
	Mayo			0.719827	1.047098		0.883463	0.88968
	Jun.			0.247686		0.365362	0.306524	0.308681
	Jul.	0.170234	0.08775				0.128992	0.129899
	Ago.	0.042961	0.031139				0.03705	0.037311

(*continúa*)

Tabla 16.19 (*continuación*)

Año	Mes	Cociente entre datos (gruesas) y promedio móvil					Promedio	Promedio ajustado a 12 (índices estacionales)
		2007	2008	2009	2010	2011		
2007	Sep.	0.316853	0.428145				0.372499	0.37512
	Oct.	1.645949		1.575791			1.61087	1.622206
	Nov.		1.394207	1.181315			1.287761	1.296823
	Dic.		1.60054	1.617021			1.608781	1.620102
						Sumas	11.91615	12

Finalmente, como se debe asegurar que la suma de estos promedios sea igual a 12, se les ajusta. Para hacer esto se obtiene su suma, la suma de todos los promedios de la penúltima columna de esta tabla 16.19, que resultó ser el 11.91615 que aparece en el último renglón de esa penúltima columna.

El ajuste consiste en multiplicar cada uno de esos promedios por el resultado de dividir 12 entre 11.91615, que es 1.007037. Así, multiplicando todos estos promedios por este factor, se obtienen los promedios ajustados de la última columna de esa tabla 16.19, cuya suma da, precisamente, 12. Estos promedios ajustados constituyen los índices estacionales para el ejemplo 16.8.

En las 2 subsecciones restantes de esta parte se ejemplifican los procedimientos que se pueden seguir para *a*) desestacionalizar series de tiempo y *b*) hacer pronósticos con índices estacionales.

16.4.2 Desestacionalización de series de tiempo

Se ilustra el procedimiento con el ejemplo siguiente.

Ejemplo 16.9

Utilizar los índices estacionales calculados en el ejemplo 16.8 para desestacionalizar los datos de ventas de gruesas de naranjas para 2007.

Solución: El procedimiento es ahora muy sencillo. Simplemente se dividen los datos originales entre su correspondiente índice estacional. En la tabla 16.20 se muestran los datos, los índices estacionales y los datos de ventas desestacionalizados.

Tabla 16.20 Datos y cálculos del ejemplo 16.9

Año	Mes	Ventas (gruesas de naranjas) 2007	Promedio ajustado a 12 (índices estacionales)	Ventas desestacionalizadas
2007	Ene.	1 920	1.323884	1 450.278
	Feb.	2 080	1.349839	1 540.925
	Mar.	2 488	1.685049	1 476.515
	Abr.	2 328	1.361406	1 709.997
	Mayo	1 624	0.88968	1 825.375
	Jun.	514	0.308681	1 665.149
	Jul.	232	0.129899	1 786.003
	Ago.	56	0.037311	1 500.898
	Sep.	392	0.37512	1 044.999
	Oct.	1 920	1.622206	1 183.573
	Nov.	1 680	1.296823	1 295.474
	Dic.	1 440	1.620102	888.8329

En la figura 16.14 se muestran las gráficas tanto de las ventas en unidades originales como las desestacionalizadas. Se observa que las ventas desestacionalizadas, la de curvas menos pronunciadas, la línea "suavizada" (que es lo que hacen los promedios móviles), muestra, precisamente, cambios menos bruscos, a la vez que mantiene la tendencia a la baja que muestran las ventas en el año.

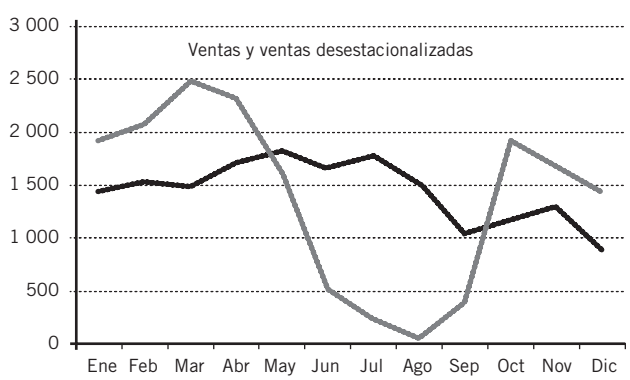


Figura 16.14 Gráficas de ventas en unidades originales y con datos desestacionalizados.

Estos datos desestacionalizados se pueden utilizar, por ejemplo y entre otras posibilidades, para mejorar la administración de inventarios, en especial para productos no perecederos.

16.4.3 Pronósticos con índices estacionales

Se pueden utilizar los índices estacionales para hacer pronósticos de ventas que tomen en cuenta las variaciones estacionales.

EJEMPLO 16.10

Si se estima que las ventas de naranjas para 2012 serán de 20 000 gruesas, pronosticar las ventas mensuales.

Solución: Sin la información de las variaciones estacionales, el pronóstico de las ventas sería el simple resultado de dividir 20 000/12 = 1 666.67. Sin embargo, si se cuenta con los índices estacionales, se puede hacer un mejor pronóstico multiplicando los índices estacionales por este promedio simple de ventas mensuales, 1 666.67. Se muestran los resultados en la tabla 16.21.

Tabla 16.21

Mes	Índices estacionales	Pronóstico de ventas
Ene.	1.323884	2 206.478
Feb.	1.349839	2 249.736
Mar.	1.685049	2 808.421
Abr.	1.361406	2 269.015
Mayo	0.88968	1 482.803
Jun.	0.308681	514.4694
Jul.	0.129899	216.4988
Ago.	0.037311	62.18512
Sep.	0.37512	625.2013
Oct.	1.622206	2 703.682
Nov.	1.296823	2 161.376
Dic.	1.620102	2 700.175
		20 000.04

Así, el pronóstico para enero se obtuvo multiplicando 1.323884 por 1 666.67. Es claro que, dada la marcada estacionalidad en las ventas de naranjas, estos pronósticos, ahora "estacionalizados" son mejores que el promedio simple.

EJERCICIOS 16.4 Variaciones estacionales

1. En la tabla siguiente se muestran los precios mensuales internacionales del oro 99.5% puro, promedio del precio diario de Londres, en dólares estadounidenses por onza troy, con datos del Banco Mundial, para todos los meses, de 2007 a 2010.

	2007	2008	2009	2010
Ene.	631.17	889.60	858.69	1 117.96
Feb.	664.75	922.30	943.00	1 095.41
Mar.	654.90	968.43	924.27	1 113.34
Abr.	679.37	909.71	890.20	1 148.69
Mayo	667.31	888.66	928.65	1 205.43
Jun.	655.66	889.49	945.67	1 232.92
Jul.	665.38	939.77	934.23	1 192.97
Ago.	665.41	839.03	949.38	1 215.81
Sep.	712.65	829.93	996.59	1 270.98
Oct.	754.60	806.62	1 043.16	1 342.02
Nov.	806.25	760.86	1 127.04	1 369.89
Dic.	803.20	816.09	1 134.72	1 390.55

Fuente: Index Mundi, "Oro Precio Mensual - Dólares americanos por onza troy", disponible en: http://www.indexmundi.com/es/precios-de-mercado/?mercancia=oro&meses=120, consultado el 26 de febrero de 2011.

a) Calcule los índices estacionales para los precios mensuales.
b) Desestacionalice las mediciones originales.
c) Calcule una ecuación de regresión lineal para los precios mensuales del oro, utilizando solamente los datos de 2009 y 2010.
d) Con la ecuación obtenida en el inciso *c*), pronostique los precios mensuales para los meses de 2011.
e) Utilice los índices estacionales obtenidos en el inciso *a*) para ajustar los pronósticos obtenidos en *d*).
f) Revise los pronósticos ajustados en el inciso *e*) contra los precios reales, los cuales se pueden consultar en el sitio que se cita como fuente de los datos de la tabla.

2. En México se utiliza el Índice Metropolitano de la Calidad del Aire (IMECA) para medir la contaminación. Los datos siguientes son mediciones cada hora, a partir de las 7 a.m. en tres días consecutivos en una zona metropolitana:

Martes	Miércoles	Jueves
62.50	70.00	87.50
70.00	75.00	105.00
87.50	87.50	112.50
125.00	120.00	175.00
150.00	150.00	180.00

(*continúa*)

(*continuación*)

Martes	Miércoles	Jueves
150.00	162.50	187.50
100.00	125.00	150.00
87.50	100.00	112.50
75.00	87.50	100.00
62.50	62.50	62.50
62.50	50.00	62.50
50.00	50.00	62.50

a) Calcule los índices estacionales para cada una de las 12 lecturas horarias.
b) Desestacionalice las mediciones originales.
c) Calcule una ecuación de regresión para los índices estacionales.
d) Con la ecuación obtenida en el inciso *c*), pronostique las lecturas de los IMECAS para las mismas horas del viernes siguiente.
e) Utilice los índices estacionales obtenidos en el inciso *a*) para ajustar los pronósticos obtenidos en *d*).

3. Una distribuidora de computadoras analiza sus ventas trimestrales y obtiene los datos que se muestran a continuación y que presentan variaciones estacionales:

Año	Trimestre	Ventas
2007	1	72
	2	62
	3	90
	4	97
2008	1	87
	2	78
	3	102
	4	111
2009	1	90
	2	84
	3	112
	4	117
2010	1	94
	2	88
	3	120
	4	126

a) Calcule los índices estacionales para las ventas trimestrales.
b) Desestacionalice las mediciones originales.
c) Calcule una ecuación de regresión para las ventas trimestrales.
d) Con la ecuación obtenida en el inciso *c*), pronostique las ventas trimestrales para los trimestres de 2011.
e) Utilice los índices estacionales obtenidos en el inciso *a*) para ajustar los pronósticos obtenidos en *d*).

16.5 Variaciones cíclicas

La medición de la tendencia, como ya se comentó, se refiere al comportamiento a mediano plazo de una serie de tiempo y es, cuando menos, de un año. Por su parte, las variaciones estacionales son las que se manifiestan en periodos inferiores a un año.

Ahora, las variaciones cíclicas son las que se refieren a variaciones en plazos de varios años, que se repiten con aproximada regularidad. Ya se mencionó en la introducción que existe el importante ciclo de negocios, o ciclo económico, que se compone de periodos sucesivos de prosperidad, recesión, depresión y recuperación y que es el resultado de variables que no tienen que ver con el clima, las costumbres sociales y las otras variables que dan cuenta de las variaciones estacionales. El estudio de los ciclos es un tema de gran importancia en sí mismo y dio lugar a gran cantidad de estudios especializados.

Existen diversas series de tiempo que, históricamente, mostraron estar relacionadas con el ciclo de negocios y que incluyen tasas de empleo y desempleo, comportamiento de tasas de interés y tipos de cambio, movimientos históricos del Producto Interno Bruto (PIB), niveles de reservas de moneda extranjera en el banco central (el Banco de México, en el caso de este país) y los índices bursátiles, como el Índice de Precios y Cotizaciones de la Bolsa Mexicana de Valores.

A estos indicadores se les clasifica como indicadores líderes, indicadores concurrentes e indicadores rezagados, según su comportamiento se adelanta, coincide o se retrasa con respecto a los cambios en los ciclos de negocios.

El análisis del componente cíclico de una serie de tiempo se debe hacer con series de tiempo con periodicidad cuando menos anual, ya que periodos menores a un año incluyen las variaciones estacionales e irregulares. Por esto, los valores anuales de una serie de tiempo representan solamente los componentes cíclico y de tendencia de la serie, con lo que el modelo se reduce a:

$$Y = T \cdot C$$

Si se despeja el componente cíclico en este modelo se obtiene:

$$C = \frac{Y}{T}$$

A partir de este modelo, multiplicando estos cocientes por 100 se obtienen los índices cíclicos. Es decir, se obtienen los índices cíclicos dividiendo los valores de la serie de tiempo entre el componente de tendencia y multiplicando el cociente por 100. En el ejemplo siguiente se ilustra el procedimiento para calcular relativos cíclicos.

Ejemplo 16.11

En la tabla 16.22 se anotan los datos del Producto Interno Bruto, PIB, con base 2003, de 1899 a 2010, para un total de 112 años.

Tabla 16.22 PIB anual, base 2003

1899	199.404	**1927**	309.19	**1955**	910.103	**1983**	4 561.93
1900	199.803	**1928**	310.318	**1956**	972.657	**1984**	4 717.53
1901	216.78	**1929**	299.28	**1957**	1 046.25	**1985**	4 820.73
1902	201.109	**1930**	279.655	**1958**	1 101.46	**1986**	4 672.31
1903	223.309	**1931**	289.467	**1959**	1 134.56	**1987**	4 752.78
1904	227.227	**1932**	246.537	**1960**	1 226.55	**1988**	4 813.77
1905	250.733	**1933**	273.521	**1961**	1 279.53	**1989**	5 011.40
1906	248.121	**1934**	291.92	**1962**	1 336.58	**1990**	5 270.78
1907	262.486	**1935**	313.998	**1963**	1 437.39	**1991**	5 492.93
1908	272.914	**1936**	339.755	**1964**	1 595.67	**1992**	5 687.44
1909	270.322	**1937**	350.795	**1965**	1 693.78	**1993**	5 797.85
1910	272.933	**1938**	355.701	**1966**	1 797.04	**1994**	6 056.55
1911	270.477	**1939**	375.326	**1967**	1 902.25	**1995**	5 679.68
1912	261.01	**1940**	380.232	**1968**	2 081.51	**1996**	5 971.54
1913	259.444	**1941**	417.029	**1969**	2 152.66	**1997**	6 376.55
1914	233.499	**1942**	441.56	**1970**	2 292.64	**1998**	6 688.32
1915	233.733	**1943**	457.504	**1971**	2 378.90	**1999**	6 947.81
1916	241.68	**1944**	494.301	**1972**	2 574.66	**2000**	7 406.51
1917	248.205	**1945**	510.247	**1973**	2 777.05	**2001**	7 394.06
1918	256.892	**1946**	543.364	**1974**	2 937.48	**2002**	7 455.36
1919	265.37	**1947**	562.989	**1975**	3 106.22	**2003**	7 555.80
1920	276.25	**1948**	585.067	**1976**	3 243.44	**2004**	7 857.72
1921	275.974	**1949**	618.184	**1977**	3 353.41	**2005**	8 103.68
1922	282.107	**1950**	678.284	**1978**	3 653.77	**2006**	8 501.26
1923	291.92	**1951**	731.027	**1979**	4 008.12	**2007**	8 810.14
1924	287.014	**1952**	760.464	**1980**	4 378.20	**2008**	8 942.35
1925	305.412	**1953**	762.916	**1981**	4 751.47	**2009**	8 398.75
1926	322.584	**1954**	838.964	**1982**	4 726.72	**2010**	8 860.70

Fuente: Aguirre Botello, Manuel, "Súper tabla bicentenario. México 1810-2011", disponible en: http://www.mexicomaxico.org/Voto/super.htm/ consultado el 18 de octubre de 2011.

Solución: En la figura 16.15 se muestra la gráfica de esos datos del PIB.

Se ajusta una función exponencial a esta serie. En la tabla 16.23 se muestran los datos y los cálculos.

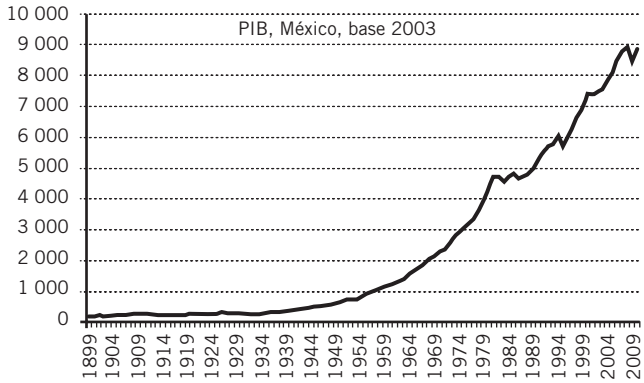


Figura 16.15 PIB anual de México, base 2003.

Tabla 16.23 Datos y cálculos para el ejemplo de PIB y variaciones cíclicas

Año (1)	PIB (2)	ln y (3)	ln $\hat{y}$ (4)	$\hat{y}$ (5)	Índice cíclico (6)
1899	199.404	5.29533292	4.76508222	117.340393	169.936366
1900	199.803	5.29733188	4.8052416	122.148608	163.573701
1901	216.78	5.37888301	4.84540098	127.153848	170.486385
1902	201.109	5.30384705	4.88556036	132.364186	151.936113
1903	223.309	5.40855646	4.92571974	137.788027	162.067057
1904	227.227	5.42594952	4.96587912	143.434117	158.419074
1905	250.733	5.52438863	5.0060385	149.311566	167.92604
1906	248.121	5.51391653	5.04619788	155.429853	159.635357
1907	262.486	5.57019775	5.08635726	161.798847	162.229834
1908	272.914	5.60915673	5.12651664	168.42882	162.035214
1909	270.322	5.59961384	5.16667602	175.330468	154.178565
1910	272.933	5.60922634	5.2068354	182.514922	149.540101
1911	270.477	5.60018707	5.24699478	189.993771	142.360983
1912	261.01	5.56455872	5.28715416	197.779078	131.970481
1913	259.444	5.55854088	5.32731354	205.883401	126.015016
1914	233.499	5.45317779	5.36747292	214.319812	108.948864
1915	233.733	5.45417944	5.4076323	223.101918	104.765123
1916	241.68	5.48761454	5.44779168	232.243886	104.063019
1917	248.205	5.51425502	5.48795106	241.760461	102.665671
1918	256.892	5.54865576	5.52811044	251.666994	102.076159
1919	265.37	5.58112508	5.56826982	261.979462	101.2942
1920	276.25	5.62130625	5.6084292	272.714501	101.29641
1921	275.974	5.62030666	5.64858858	283.889426	97.2117925
1922	282.107	5.64228643	5.68874796	295.522262	95.4604901
1923	291.92	5.67647979	5.72890734	307.631773	94.8926691
1924	287.014	5.659531	5.76906672	320.23749	89.6253589
1925	305.412	5.72166168	5.8092261	333.359748	91.61634
1926	322.584	5.77636357	5.84938548	347.019711	92.9584083
1927	309.19	5.73395597	5.88954486	361.239414	85.5914354
1928	310.318	5.73759758	5.92970424	376.041793	82.5222105
1929	299.28	5.70137959	5.96986362	391.450724	76.4540673
1930	279.655	5.6335567	6.010023	407.491061	68.6284994
1931	289.467	5.6680413	6.05018238	424.188677	68.2401525

Año (1)	PIB (2)	ln y (3)	ln $\hat{y}$ (4)	$\hat{y}$ (5)	Índice cíclico (6)
1932	246.537	5.50751208	6.09034176	441.570505	55.8318541
1933	273.521	5.6113784	6.13050114	459.664582	59.5044758
1934	291.92	5.67647979	6.17066052	478.500094	61.0073025
1935	313.998	5.74938662	6.2108199	498.107422	63.0382095
1936	339.755	5.82822477	6.25097928	518.518192	65.5242198
1937	350.795	5.86020201	6.29113866	539.765327	64.9902804
1938	355.701	5.87409049	6.33129804	561.883099	63.3051609
1939	375.326	5.92779498	6.37145742	584.907182	64.1684717
1940	380.232	5.94078159	6.4116168	608.874716	62.4483149
1941	417.029	6.03315576	6.45177618	633.824358	65.7956727
1942	441.56	6.09031391	6.49193556	659.796353	66.92368
1943	457.504	6.12578563	6.53209494	686.832593	66.6107003
1944	494.301	6.20314464	6.57225432	714.976687	69.1352612
1945	510.247	6.23489492	6.6124137	744.274032	68.5563351
1946	543.364	6.29777945	6.65257308	774.771884	70.1321268
1947	562.989	6.33326009	6.69273246	806.519436	69.8047654
1948	585.067	6.37172637	6.73289184	839.567895	69.6866809
1949	618.184	6.42678615	6.77305122	873.97057	70.7328166
1950	678.284	6.51956608	6.8132106	909.78295	74.5544857
1951	731.027	6.59445039	6.85336998	947.062802	77.1888621
1952	760.464	6.63392877	6.89352936	985.870257	77.1363163
1953	762.916	6.63714793	6.93368874	1 026.26791	74.3388731
1954	838.964	6.7321678	6.97384812	1 068.32093	78.5310836
1955	910.103	6.81355778	7.0140075	1 112.09713	81.8366466
1956	972.657	6.8800315	7.05416688	1 157.66714	84.0187102
1957	1 046.25	6.95296858	7.09432626	1 205.10445	86.8182834
1958	1 101.46	7.00438731	7.13448564	1 254.48559	87.8013274
1959	1 134.56	7.03400283	7.17464502	1 305.8902	86.8804285
1960	1 226.55	7.11196389	7.2148044	1 359.4012	90.227521
1961	1 279.53	7.15424967	7.25496378	1 415.10489	90.4195869
1962	1 336.58	7.19787164	7.29512316	1 473.09114	90.7332181
1963	1 437.39	7.27058425	7.33528254	1 533.45347	93.7354816
1964	1 595.67	7.37504648	7.37544192	1 596.28925	99.9609562
1965	1 693.78	7.434718	7.4156013	1 661.69983	101.930564
1966	1 797.04	7.49389392	7.45576068	1 729.79071	103.887481
1967	1 902.25	7.5507932	7.49592006	1 800.67173	105.641188
1968	2 081.51	7.64084647	7.53607944	1 874.45721	111.045746
1969	2 152.66	7.67446142	7.57623882	1 951.26618	110.321392
1970	2 292.64	7.73745971	7.6163982	2 031.22251	112.870007
1971	2 378.90	7.77439347	7.65655758	2 114.4552	112.506522
1972	2 574.66	7.85347121	7.69671696	2 201.09847	116.971414
1973	2 777.05	7.92914521	7.73687634	2 291.2921	121.200261
1974	2 937.48	7.98530667	7.77703572	2 385.18156	123.155321
1975	3 106.22	8.04116215	7.8171951	2 482.9183	125.103633
1976	3 243.44	8.08438854	7.85735448	2 584.65995	125.487919
1977	3 353.41	8.11773302	7.89751386	2 690.57064	124.635642

(*continúa*)

Tabla 16.23 (*continuación*)

Año (1)	PIB (2)	ln y (3)	ln $\hat{y}$ (4)	$\hat{y}$ (5)	Índice cíclico (6)
1978	3 653.77	8.20351561	7.93767324	2 800.8212	130.453633
1979	4 008.12	8.29607808	7.97783262	2 915.58945	137.472098
1980	4 378.20	8.38439342	8.017992	3 035.06052	144.254191
1981	4 751.47	8.46620932	8.05815138	3 159.42712	150.390239
1982	4 726.72	8.46098764	8.09831076	3 288.88983	143.717918
1983	4 561.93	8.42550106	8.13847014	3 423.6575	133.247265
1984	4 717.53	8.45904042	8.17862952	3 563.9475	132.368084
1985	4 820.73	8.48068148	8.2187889	3 709.9861	129.939409
1986	4 672.31	8.44940802	8.25894828	3 862.00888	120.981234
1987	4 752.78	8.46648583	8.29910766	4 020.26104	118.220781
1988	4 813.77	8.47923667	8.33926704	4 184.99785	115.024528
1989	5 011.40	8.51947139	8.37942642	4 356.48501	115.033197
1990	5 270.78	8.56993421	8.4195858	4 534.99915	116.224564
1991	5 492.93	8.61121782	8.45974518	4 720.82819	116.355304
1992	5 687.44	8.64601622	8.49990456	4 914.27189	115.733198
1993	5 797.85	8.66524175	8.54006394	5 115.64225	113.335642
1994	6 056.55	8.70889511	8.58022332	5 325.2641	113.732331
1995	5 679.68	8.64465052	8.6203827	5 543.47555	102.457059
1996	5 971.54	8.69476013	8.66054208	5 770.62858	103.481621
1997	6 376.55	8.76038248	8.70070146	6 007.08957	106.150407
1998	6 688.32	8.80811815	8.74086084	6 253.23994	106.95769
1999	6 947.81	8.84618121	8.78102022	6 509.47672	106.73371
2000	7 406.51	8.91011476	8.8211796	6 776.21323	109.301622
2001	7 394.06	8.90843225	8.86133898	7 053.87971	104.822598
2002	7 455.36	8.91668838	8.90149836	7 342.92402	101.531202
2003	7 555.80	8.93007129	8.94165774	7 643.8124	98.8486321
2004	7 857.72	8.96925177	8.98181712	7 957.03018	98.751919
2005	8 103.68	9.00007356	9.0219765	8 283.08257	97.8341086
2006	8 501.26	9.04796943	9.06213588	8 622.49549	98.5939396
2007	8 810.14	9.08365861	9.10229526	8 975.81642	98.1541911
2008	8 942.35	9.0985537	9.14245464	9 343.61525	95.7054605
2009	8 398.75	9.03583816	9.18261402	9 726.48525	86.3492802
2010	8 860.70	9.08938139	9.2227734	10 125.044	87.512736

Las columnas (1) y (2) son los datos, la (3) son los logaritmos del PIB. Con los datos de las columnas (1) y (2) se corrió la opción "Regresión" de Excel y se obtuvieron los parámetros de la ecuación de regresión lineal, los cuales condujeron a la ecuación:

$$\ln \hat{y} = -71.4975804 + 0.04015938x$$

Con esta ecuación de regresión se calcularon los datos de la columna (4). La columna (5) se obtuvo con los antilogaritmos de los datos de la columna (4).

Ahora, en la figura 16.16 se muestra la gráfica de los datos originales de PIB, junto con la función exponencial ajustada, $\hat{y}$. Se puede apreciar en la gráfica que el ajuste parece ser bastante cercano pero, de nuevo, su evaluación estadística está fuera del alcance de este libro.

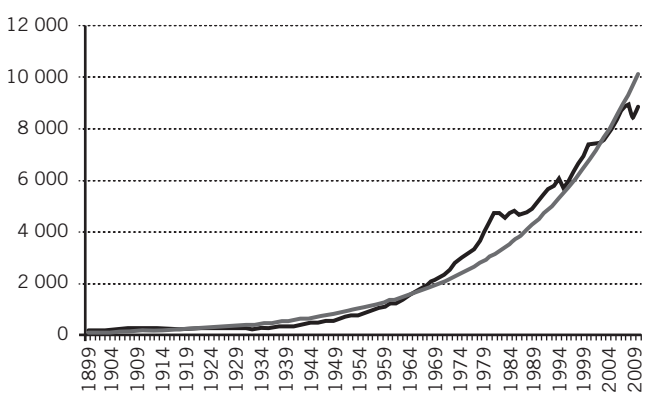


Figura 16.16 IPC y curva exponencial ajustada con mínimos cuadrados.

Finalmente, en la columna (6) de la tabla 16.23 se calcularon los índices cíclicos, dividiendo los datos originales de PIB entre la función ajustada ($\hat{y}$) y multiplicando este cociente por 100. En la figura 16.17 se grafican estos índices cíclicos y se puede apreciar que parece existir un ciclo, no regular, con una cima de, aproximadamente, 1989 y hasta la cima de 1932 (33 años), seguida de otra cima en 1981 (49 años), seguida ésta, a su vez, por una racha a la baja que parece que aún no termina.

Igual que antes, un análisis más cuidadoso de esta información requiere de textos especializados de series de tiempo o de econometría.

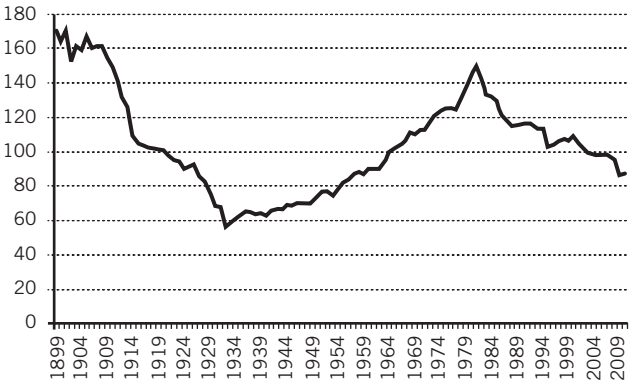


Figura 16.17 Índices cíclicos del PIB mexicano.

ejercicios 16.5 Variaciones cíclicas

1. Se presentan en seguida los datos de ventas anuales de cierto producto.

Año	Ventas (miles de pesos)
1979	11.58
1980	10.86
1981	11.94
1982	12.56
1983	13.07
1984	14.31
1985	14.01
1986	13.36
1987	14.40
1988	14.45
1989	13.39
1990	14.83
1991	15.41
1992	16.35
1993	13.70
1994	13.17
1995	14.75
1996	15.83
1997	15.95
1998	15.36
1999	13.76
2000	13.44
2001	12.75
2002	13.92
2003	15.12
2004	15.89
2005	16.14
2006	15.18
2007	15.40
2008	14.85
2009	14.10
2010	13.23

a) Grafique los datos.
b) Calcule los índices cíclicos y grafíquelos.
c) Interprete los índices cíclicos.

2. Con los datos de la deuda pública mexicana en pesos corrientes:

a) Grafique los datos.
b) Calcule los índices cíclicos y grafíquelos.
c) Interprete los índices cíclicos.

1993+	256 716.52
1994	291 336.08
1995	666 162.42
1996	751 876.43
1997	709 219.24
1998	917 407.33
1999	878 596.04
2000	791 857.87
2001	736 708.63
2002	816 560.70
2003	885 063.20
2004	888 913.48
2005	767 639.25
2006	595 854.08
2007	603 368.41
2008	784 050.03

(*continúa*)

(*continuación*)

2009	1 256 443.12
2010	1 367 098.64

Fuente: Aguirre Botello, Manuel, "Súper tabla bicentenario. México 1810-2011", disponible en: http://www.mexicomaxico.org/Voto/super.htm, consultado el 18 de octubre de 2011.

3. Los datos siguientes son la inflación anual en México.

Año	Inflación %
2001	4.4
2002	5.7
2003	3.98
2004	5.19
2005	3.33
2006	4.05
2007	3.76
2008	6.53
2009	3.57
2010	4.4

Fuente: Aguirre Botello, Manuel, "Súper tabla bicentenario. México 1810-2011", disponible en: http://www.mexicomaxico.org/Voto/super.htm, consultado el 25 de febrero de 2012.

a) Grafique los datos.
b) Calcule los índices cíclicos y grafíquelos.
c) Interprete los índices cíclicos.

16.6 Resumen

Se describen en este capítulo las principales técnicas que se utilizan para modelar series de tiempo, que son conjuntos de observaciones que se generan con el tiempo y se explicó que existen 2 modelos principales, el aditivo y el multiplicativo y se asentó que el que más comúnmente se utiliza es este último:

$$Y = T \cdot E \cdot C \cdot I$$

En donde esta expresión contiene los 4 componentes principales de una serie de tiempo: tendencia secular, variaciones estacionales, variaciones cíclicas y variaciones irregulares.

Se ilustró la forma en la que se puede modelar el componente de tendencia mediante promedios móviles exponenciales y mediante el ajuste de una recta, una función exponencial o mediante una parábola.

Se presentaron ejemplos de cómo se puede aislar el componente estacional y cómo se le utiliza para "desestacionalizar" series de tiempo, es decir, para eliminar en la serie este componente estacional.

Se ejemplificaron también los procedimientos que se utilizan para construir índices cíclicos y que consisten, básicamente, en eliminar el elemento de tendencia de la serie ya que, como se trata de datos anuales, ya no está presente el componente estacional. Así, al eliminar el componente de tendencia, lo que queda es ese componente cíclico.

16.7 Fórmulas del capítulo

16.1 Modelo clásico de series de tiempo

El modelo aditivo de series de tiempo:

$$Y = T + C + E + I \qquad (16.1)$$

El modelo multiplicativo de series de tiempo:

$$Y = T \cdot E \cdot C \cdot I \qquad (16.2)$$

16.3 Suavización con promedios móviles exponenciales

Promedios móviles exponenciales:

$$PME_t = PME_{t-1} + w\left(Y_t - PME_{t-1}\right) \qquad (16.3)$$

16.8 Ejercicios adicionales

16.3 Tendencia secular
Ajuste y pronósticos con promedios móviles

1. Ingrese a la sección de finanzas de Yahoo! (http://mx.finance.yahoo.com) en internet y obtenga datos históricos para los precios de alguna de las acciones que cotizan en la Bolsa Mexicana de Valores (BMV). Algunos ejemplos de las claves que se pueden anotar en el globo de "Buscar cotizaciones" de esa página de Yahoo! para encontrar cotizaciones de acciones mexicanas que cotizan en esa BMV son (a todos ellos se les debe añadir ".mx" al final):

ALFAA	CHDRAUIB	GMODELOC	OHLMEX
AMXL	COMERCIUBC	GRUMAB	TELMEXL

ARA*	COMPARC	HOMEX	TLEVISACPO
ASURB	ELEKTRA	ICA*	URBI
AUTLANB	FEMSAUBD	KIMBERA	WALMEXV
AZTECACPO	GAPB	KOFL	
BIMBOA	GEOB	LABB	
BOLSAA	GFINBURO	MEXCHEM	
C	GFNORTEO	MFRISCOA1	
CEMEXCPO	GMEXICOB	NAFTRAC02	

a) Ingrese a la sección de "Precios históricos" cuya liga está del lado izquierdo de la página de cada acción y obtenga cuando menos 100 de las observaciones más recientes.

b) Obtenga promedios móviles de los precios de cierre con periodo de 5 y de 20 días y grafíquelos junto con los datos de los precios de cierre.

c) Observe la gráfica y describa cómo se comportan los precios de cierre cuando los 2 promedios móviles se cruzan.

2. En la tabla siguiente se anota el tipo de cambio promedio peso-dólar estadounidense de 1993 a 2010.

1993	3.26
1994	3.41
1995	6.6
1996	7.65
1997	8.03
1998	9.94
1999	9.52
2000	9.36
2001	9.17
2002	10.36
2003	11.2
2004	11.22
2005	10.71
2006	10.88
2007	10.9
2008	13.77
2009	13.04
2010	12.38

Fuente: Aguirre Botello, Manuel, "Súper tabla bicentenario. México 1810-2011", disponible en http://www.mexicomaxico.org/Voto/super.htm, consultado el 25 de febrero de 2012.

Ensaye ajustes con una serie de promedios móviles exponenciales con diferentes factores de suavización hasta encontrar una que resulte satisfactoria y explicar cómo se llegó a ésta.

3. Los datos siguientes corresponden a montos de cartera bancaria vencida, en millones de pesos.

Mes	2006	2007	2008	2009	2010
Ene.	25 977.80	28 738.80	44 280.10	63 634.60	60 242.50
Feb.	26 594.40	30 333.40	44 397.50	64 963.10	58 123.20
Mar.	21 496.00	31 554.70	44 397.50	64 343.20	54 747.60
Abr.	21 804.70	32 869.70	49 427.20	67 683.00	56 315.90
Mayo	23 151.80	34 079.10	51 856.50	71 359.40	57 863.20
Jun.	24 031.20	35 068.00	50 969.40	71 123.90	54 122.10
Jul.	24 777.30	38 018.60	53 007.40	69 973.30	55 066.70
Ago.	25 520.10	38 906.00	54 445.60	66 622.50	53 561.10
Sept.	26 421.50	40 383.60	56 860.30	64 702.00	51 495.60
Oct.	27 456.10	41 859.30	62 982.70	64 476.30	51 636.90
Nov.	29 032.10	42 603.20	66 861.00	64 615.80	52 344.20
Dic.	27 350.10	43 086.50	60 287.30	60 829.70	49 987.40

Fuente: INEGI, "Banco de información económica", disponible en: http://dgcnesyp.inegi.gob.mx/cgi-win/bdieintsi.exe/Consultar/, consultado el 10 de junio de 2011.

Ensaye ajustes con una serie de promedios móviles exponenciales con diferentes factores de suavización hasta encontrar una que resulte satisfactoria y explicar cómo se llegó a ésta.

16.3.2 Ajuste de rectas y curvas con mínimos cuadrados Recta ascendente

4. Ajuste a los siguientes datos sobre el gasto educativo total en educación en el sistema escolarizado, por el método de mínimos cuadrados la curva o la recta más apropiada y, con base en su ecuación, estime cuál será el gasto total en educación para 2015.

Año	Gasto educativo total en educación en el sistema escolarizado
1995	85 858.4
1996	122 947.1
1997	155 889.5
1998	192 124.1
1999	227 910.3
2000	276 435.6
2001	311 174.7
2002	344 332.1
2003	386 715.7
2004	416 161.1
2005	464 030.1
2006	503 724.2
2007	543 583.8
2008 R/	600 985.9
2009	636 178.3
2010 E/	656 271.2

Fuente: INEGI, "Anuario estadístico de los Estados Unidos Mexicanos", disponible en: www.inegi.org.mx, consultado el 22 de febrero de 2012.

16.3.3 Ajuste de una función exponencial con mínimos cuadrados

5. Ingrese al sitio del Banco de México en internet (www.banxico.gob.mx) y continúe con las siguientes ligas:

- Estadísticas.
- Sistema financiero
- Mercados financieros (tipo de cambio, tasas de interés y derivados).

En la sección de Mercado cambiario revise las secciones:

- Tipos de cambio.
- Cotización de las divisas que conforman la canasta de DEG.

Dé clic sobre el ícono "XLS" que aparece en la parte superior izquierda de la página; entonces se abre un archivo de Excel con varias columnas que contiene datos sobre, precisamente, cotizaciones con respecto al peso mexicano de dólar estadounidense, yen japonés, libra esterlina, euros y los propios DEG.

a) Con los datos completos de la cotización del peso respecto al euro, o alguna parte apropiada o de interés, ajuste por el método de mínimos cuadrados una curva exponencial.
b) Con los datos completos de la cotización del peso respecto al dólar, o alguna parte apropiada o de interés, ajuste por el método de mínimos cuadrados la curva o la recta más apropiada.
c) Con los datos de las otras columnas, ajuste por el método de mínimos cuadrados la curva o la recta más apropiada.

Recta ascendente

6. Ajuste a los siguientes datos sobre la esperanza de vida al nacer, por el método de mínimos cuadrados la curva o la recta más apropiada y, con base en su ecuación, estime cuál será la esperanza de vida al nacer para 2015.

Año	Tasa bruta de mortalidad (por 1 000 habitantes)
1995	72.5
1996	72.8
1997	73.2
1998	73.5
1999	73.8
2000	74.0
2001	74.3
2002	74.4
2003	74.5
2004	74.5
2005	74.5
2006	74.8
2007	75.0
2008	75.1
2009	75.3
2010	75.4
2011	75.6

Fuente: INEGI, "Anuario estadístico de los Estados Unidos Mexicanos", disponible en: www.inegi.org.mx, consultado el 17 de octubre de 2011.

Exponencial

7. Los datos siguientes representan las ventas mensuales de un nuevo equipo de terapia respiratoria. Ajuste a estos datos, por el método de mínimos cuadrados, la curva o la recta más apropiada y, con base en su ecuación, estime cuáles serán las ventas para el próximo mes.

4
8
6
10
15
20
30
35
40
45
55
70
95
109

16.4 Variaciones estacionales

8. Los datos siguientes corresponden a los gastos mensuales de mantenimiento de una unidad de casas en condominio horizontal.

Mes	2008	2009	2010
Ene.	17 000	18 000	19 500
Feb.	18 000	20 500	21 000
Mar.	20 500	21 500	23 000
Abr.	23 000	24 500	28 000
Mayo	24 000	26 500	29 000
Jun.	31 500	33 000	39 000
Jul.	36 000	40 000	42 000
Ago.	29 000	33 500	33 000
Sept.	24 000	26 000	29 000
Oct.	24 000	27 000	29 500

Mes	2008	2009	2010
Nov.	23 000	25 500	28 000
Dic.	19 500	22 000	25 000

a) Calcule los índices estacionales para los gastos mensuales.
b) Desestacionalice las mediciones originales.
c) Calcule una ecuación de regresión para los índices estacionales.
d) Con la ecuación obtenida en *c*), pronostique los gastos mensuales para 2011.
e) Utilice los índices estacionales obtenidos en *a*) para ajustar los pronósticos obtenidos en *d*).

9. Los siguientes datos corresponden a las ventas trimestrales (en miles de pesos) alcanzadas por una compañía en el transcurso de 10 años.

	Ventas trimestrales (miles de pesos)			
Año	T1	T2	T3	T4
1	747	927	783	1 215
2	882	1 089	909	1 386
3	1 089	1 305	1 143	1 539
4	1 233	1 440	1 278	1 728
5	1 566	1 773	1 620	2 079
6	1 638	1 845	1 674	2 160
7	1 800	1 998	1 845	2 259
8	2 007	2 259	2 061	2 493
9	2 223	2 421	2 259	2 682
10	2 322	2 583	2 340	2 898

a) Calcule los índices estacionales para las ventas trimestrales.
b) Desestacionalice las mediciones originales.
c) Calcule una ecuación de regresión para los índices estacionales.
d) Con la ecuación obtenida en el inciso *c*), pronostique las ventas trimestrales para el año 11.
e) Utilice los índices estacionales obtenidos en el inciso *a*) para ajustar los pronósticos obtenidos en *d*).

10. Se presentan a continuación los datos sobre ventas de automóviles realizadas por una red de distribuidoras de autos nuevos, para los años 2006-2010.

	Ventas trimestrales (núm. de autos)			
Año	T1	T2	T3	T4
2006	423	383	237	145
2007	449	414	361	201
2008	508	481	410	303
2009	578	561	441	383
2010	621	625	511	483

a) Calcule los índices estacionales para las ventas trimestrales.
b) Desestacionalice las mediciones originales.
c) Calcule una ecuación de regresión para los índices estacionales.
d) Con la ecuación obtenida en el inciso *c*), pronostique las ventas trimestrales para el año 2011.
e) Utilice los índices estacionales obtenidos en el inciso *a*) para ajustar los pronósticos obtenidos en *d*).

11. Se muestran en seguida los datos del total de turistas que ingresaron al país, mes por mes, de enero de 2007 a diciembre de 2010, con datos del INEGI.

Miles de turistas que ingresaron a México				
Mes	2007	2008	2009	2010
Ene.	8 147.80	7 829.80	7 588.80	7 128.10
Feb.	7 874.20	7 762.80	7 089.80	6 543.60
Mar.	8 247.60	8 538.60	7 800.50	7 611.20
Abr.	7 821.60	7 568.30	7 349.80	7 007.70
Mayo	7 449.00	7 619.60	6 208.00	6 764.80
Jun.	7 279.00	7 495.10	6 585.20	6 481.60
Jul.	7 375.30	7 819.50	7 298.30	6 416.40
Ago.	7 338.40	7 278.60	7 004.50	6 269.30
Sept.	6 568.40	6 392.00	6 651.10	5 797.70
Oct.	7 609.00	7 222.50	7 201.60	6 157.80
Nov.	7 707.40	7 458.30	7 098.30	6 226.30
Dic.	8 761.30	8 477.20	8 313.10	7 361.40

Fuente: INEGI, "Banco de Información Económica", disponible en: http://dgcnesyp.inegi.gob.mx/cgi-win/bdieintsi.exe/Consultar, consultado el 18 de octubre de 2011.

a) Calcule los índices estacionales para el número mensual de turistas que ingresan.
b) Desestacionalice las mediciones originales.
c) Calcule una ecuación de regresión para los índices estacionales.
d) Con la ecuación obtenida en el inciso *c*), pronostique el ingreso de turistas para cada mes de 2011.
e) Utilice los índices estacionales obtenidos en el inciso *a*) para ajustar los pronósticos obtenidos en *d*).

16.5 Variaciones cíclicas

12. En la tabla siguiente se muestra la información sobre la capacidad total de energía eólica instalada en el mundo, en megavatios, según ciertas estimaciones.

Año	Capacidad total de energía eólica instalada en el mundo (megavatios)
1976	4.704
1977	3.92
1978	5.39
1979	15.288
1980	22.638
1981	22.834
1982	30.772

(*continúa*)

(*continuación*)

Año	Capacidad total de energía eólica instalada en el mundo (megavatios)
1983	45.08
1984	45.178
1985	41.062
1986	44.59
1987	52.43
1988	47.432
1989	60.368
1990	64.288
1991	69.972
1992	81.732
1993	91.728
1994	92.316
1995	83.692
1996	84.476
1997	88.102
1998	87.416
1999	97.118
2000	98.294
2001	109.466
2002	106.036
2003	113.19
2004	116.816
2005	122.696
2006	133.574
2007	143.864
2008	143.178
2009	148.372
2010	147.882

a) Grafique los datos.
b) Calcule los índices cíclicos y grafíquelos.
c) Interprete los índices cíclicos.

13. Los datos siguientes corresponden a los miles de piezas de correspondencia manejados por el Servicio Postal Mexicano en los años anotados.

Periodo	Total
1990	796 929
1991	846 996
1992	886 829
1993	936 412
1994	980 332
1995	95 143
1996	983 973
1997	1 169 123
1998	1 132 960
1999	749 804
2000	1,159 750
2001	1 017 792
2002	647 533
2003	684 470
2004	702 935
2005	731 790
2006	776 967
2007	914 393
2008	961 165

Fuente: INEGI, "Banco de Información Económica", disponible en: http://dgcnesyp.inegi.gob.mx/cgi-win/bdieintsi.exe/Consultar, consultado el 18 de octubre de 2011.

a) Grafique los datos.
b) Calcule los índices cíclicos y grafíquelos.
c) Interprete los índices cíclicos.

14. Los datos sobre saldo de exportaciones contra importaciones (FOB, *Free on Board*, es decir, con todos los gastos incluidos hasta su colocación a bordo del medio de transporte) que registra el Banco de Información Económica del INEGI para los meses de 2007 a 2010 son los siguientes:

Mes	2007	2008	2009	2010
Ene.	–1 677.50	–1 763.50	–1 707.10	–444.2
Feb.	–398.5	–1 122.90	–548.5	422.8
Mar.	–377.6	1 238.60	–36.4	393.2
Abr.	–759.3	–1 088.20	168.5	80.5
Mayo	–721.9	–12.4	449.7	172.5
Jun.	–826.7	235.6	–168.3	–336.4
Jul.	–694.8	–1 214.60	–1 238.60	–1 030.30
Ago.	–1 061.50	–2 277.50	–668.6	–697.8
Sept.	–638	–2 935.30	–879	–523
Oct.	–1 560.90	–3 292.30	236.8	–787.7
Nov.	–821	–2 959.30	–129.4	–84.7
Dic.	–535.8	–2 068.90	–160.5	–173.6

Fuente: INEGI, "Banco de Información Económica", disponible en: http://dgcnesyp.inegi.gob.mx/cgi-win/bdieintsi.exe/Consultar, consultado el 18 de octubre de 2011.

a) Grafique los datos.
b) Calcule los índices cíclicos y grafíquelos.
c) Interprete los índices cíclicos.

capítulo 17

Pruebas estadísticas no paramétricas

Sumario

17.1 Las pruebas de hipótesis no paramétricas en este libro
17.2 Prueba de rachas para aleatoriedad de Wald-Wolfowitz
- 17.2.1 Características
- 17.2.2 Excel y el cálculo de probabilidades para la prueba de rachas de Wald-Wolfowitz

17.3 Prueba de los signos
- 17.3.1 Características
- 17.3.2 Prueba del signo para una muestra pequeña
- 17.3.3 Prueba del signo para una muestra grande (aproximación normal)
- 17.3.4 Prueba del signo para 2 muestras apareadas pequeñas
- 17.3.5 Prueba del signo para 2 muestras apareadas grandes (aproximación normal)

17.4 Prueba de rangos con signo de Wilcoxon
- 17.4.1 Características
- 17.4.2 Prueba de rangos con signo de Wilcoxon para una muestra pequeña
- 17.4.3 Excel y la prueba de rangos con signo de Wilcoxon
- 17.4.4 Prueba de rangos con signo de Wilcoxon para una muestra grande (aproximación normal)
- 17.4.5 Prueba de rangos con signo de Wilcoxon para 2 muestras apareadas pequeñas
- 17.4.6 Prueba de rangos con signo de Wilcoxon para 2 muestras apareadas grandes (aproximación normal)

17.5 Prueba *U* de Mann-Whitney para 2 muestras independientes
- 17.5.1 Características

17.6 Prueba de suma de rangos de Kruskal-Wallis para más de 2 medias
17.7 Prueba de Friedman para diseños en bloques aleatorizados
17.8 El coeficiente de correlación por rangos de Spearman
17.9 Resumen
17.10 Fórmulas del capítulo
17.11 Ejercicios adicionales

En la sección 11.12 del capítulo 11, que trata de las pruebas de hipótesis que se realizan con la distribución ji cuadrada, se mencionó que a los procedimientos de pruebas de hipótesis que se revisaron hasta el capítulo 10 se les conoce como *métodos parámetricos* o *métodos clásicos* y se distinguen (sin que se haga una distinción enteramente tajante) de las pruebas no paramétricas, o métodos no paramétricos, que son el tema de este capítulo. La diferenciación entre unos y otros se hace contrastando las siguientes características que suelen tener las pruebas paramétricas:

1. Se realizan sobre valores que se supone tienen los parámetros de una o más poblaciones (por ejemplo, una prueba sobre la media de una población, μ, o sobre la diferencia entre las proporciones de 2 poblaciones, $\pi_1 - \pi_2$). Es precisamente esta característica la que les da el nombre de pruebas *parámetricas*.
2. Los datos deben medirse en una escala que sea cuando menos de intervalo. Esto permite, entre otras cosas, calcular medias y desviaciones estándar que, como se vio antes, se utilizan para realizar las pruebas de hipótesis. Aquí conviene recordar la clasificación que se hizo de las escalas de medición en la introducción del libro, sección 1.6 del primer capítulo, y que se reproduce en la tabla 17.1.

Tabla 17.1 Escalas de medición

	Categorías	Tipo
Nominal	Categorías sin orden	Cualitativa
Ordinal	Categorías con orden	Cualitativa
De intervalo	Números cero no absoluto	Cuantitativa
De razón	Números cero absoluto	Cuantitativa

3. Para algunas de estas pruebas se hacen suposiciones sobre la forma de la distribución de la variable en la población como, por ejemplo, cuando se establece que la variable se distribuirá de manera normal en la población como requisito para poder aplicar la distribución t para una muestra pequeña.

En el capítulo 11 se revisaron pruebas que se consideran tanto paramétricas como no paramétricas. Las pruebas sobre una varianza o sobre 1, 2 o n proporciones se clasificarían, de acuerdo con los 2 criterios anteriores, como pruebas paramétricas, en tanto que las demás pruebas son no paramétricas porque no se ajustan a cuando menos 1 de esos 3 criterios.

Las pruebas de bondad de ajuste no se realizan sobre ningún parámetro poblacional, la prueba consiste en evaluar si determinados datos se ajustan o no a alguna distribución teórica o empírica y las pruebas para la independencia entre 2 variables son también pruebas no paramétricas por la misma razón: se prueba precisamente la independencia entre variables y no algo relacionado con algún parámetro poblacional.

Además, por otro lado, las variables que se utilizan para hacer la doble clasificación de los datos en la tabla suelen ser variables en escala nominal u ordinal (sexo, estado civil, etc.).

Por otro lado, la prueba de χ^2 para la diferencia entre las varianzas de 2 poblaciones es, literalmente, una prueba paramétrica, ya que la hipótesis nula se plantea en términos de estos parámetros poblacionales.

Así, las pruebas no parámetricas no poseen cuando menos 1 de las 3 características anteriores; es decir:

- No se hace la prueba sobre un parámetro.
- La escala de medición de las variables es nominal u ordinal.
- No se requieren suposiciones sobre la forma de la distribución de la variable en la población.

Como se verá más adelante, existen diversas pruebas en las que los datos se miden en escala ordinal y en ellos se trabaja con medianas en vez de utilizar medias aritméticas como se hace con las pruebas paramétricas, en otras palabras, como no se pueden calcular medias aritméticas ni otras medidas con datos medidos en escala ordinal, se utiliza la mediana, que es una medida que sí es adecuada calcular para este tipo de datos.

Para concluir este tema de la diferenciación entre pruebas paramétricas y pruebas no paramétricas, es necesario comentar que la distinción entre unas y otras en términos de las 3 características que se acaban de mencionar no siempre es clara. Empezando por las de χ^2 hay pruebas que se pueden clasificar en las 2 clases con este mismo estadístico de prueba; otro ejemplo: como se verá más adelante, existen pruebas que se clasifican como no paramétricas y que son pruebas que se hacen sobre la mediana de 1 o más poblaciones. Como la mediana de una población es un parámetro poblacional, estas pruebas serían, literalmente, paramétricas. Sin embargo, como la o las variables implicadas pudieran darse en escala nominal u ordinal, se les puede clasificar también como pruebas no paramétricas. Es posible que la principal utilidad de tener presentes las 3 características mencionadas sea lograr una mejor comprensión de todas las pruebas estadísticas en su conjunto.

Un detalle que es importante tener presente a lo largo de este capítulo es que el procedimiento básico de las pruebas parámetricas consiste en calcular, a partir de los datos muestrales, un estadístico de prueba (z, t, χ^2, etc.) —al cual se le suele denominar *valor calculado del estadístico de prueba* o *valor empírico del estadístico de prueba*—, para compararlo con el valor crítico del estadístico de prueba que se determina de acuerdo al valor especificado de α; el nivel de significación. Con base en esta comparación, tomar la decisión correspondiente respecto a la hipótesis nula.

En las pruebas no paramétricas sucede lo mismo, pero muchas de ellas usan un estadístico de prueba particular cuando la muestra es pequeña y utilizan una aproximación normal, con z, cuando la muestra es grande.

En la sección siguiente se presenta un panorama general de las pruebas de hipótesis no paramétricas.

17.1 Pruebas de hipótesis no paramétricas en este libro

En el cuadro siguiente se listan las pruebas no paramétricas que se revisan en este libro, en este capítulo y en el capítulo 11, que trata de las pruebas que se realizan con la ji cuadrada como estadístico de prueba. El propósito de este cuadro es ofrecer un panorama general de este tipo de pruebas junto con sus principales características con la intención principal de facilitar su comprensión y su utilización.

El cuadro se divide en 5 secciones:

1. Pruebas para una muestra.
2. Pruebas para 2 muestras independientes.

3. Pruebas para 2 muestras relacionadas.
4. Pruebas para más de 2 muestras independientes.
5. Pruebas para más de 2 muestras relacionadas.

Se anota para cada prueba el motivo de la prueba, es decir, sobre qué se plantea la hipótesis nula, el nivel de medición de la o las variables involucradas, el nombre y el capítulo y la sección en donde se revisa la prueba.

Pruebas para una muestra			
Tipo de prueba	**Nivel mínimo de escala de medición de la variable**	**Prueba**	**Secciones del libro en que se estudia**
Bondad de ajuste	Cualquiera	χ^x, ji cuadrada de bondad de ajuste a distribuciones teóricas	11.9
Bondad de ajuste	Cualquiera	χ^x, ji cuadrada de bondad de ajuste a distribuciones empíricas	11.10
Aleatoriedad		Rachas para aleatoriedad de Wald-Wolfowitz	17.2
Mediana	Ordinal	De los signos	17.3
Mediana	Intervalo	Rangos con signo de Wilcoxon	17.4
Pruebas para 2 muestras independientes			
Tipo de prueba	**Nivel mínimo de escala de medición de la variable**	**Prueba**	**Secciones del libro en que se estudia**
Independencia entre 2 variables	De intervalo	χ^x, ji cuadrada de independencia	11.11
Igualdad de medias	De intervalo	De Mann-Whitney	17.5
Igualdad de medianas	Ordinal	Suma de rangos de Wilcoxon	17.4
Correlación	Ordinal	El coeficiente de correlación por rangos de Spearman	17.8
Pruebas para 2 muestras relacionadas (pareadas)			
Tipo de prueba	**Nivel mínimo de escala de medición de la variable**	**Prueba**	**Secciones del libro en que se estudia**
Igualdad de medianas	Ordinal	De los signos	17.2
Igualdad de medianas	Intervalo	Rangos con signo de Wilcoxon	17.4
Pruebas para más de 2 muestras independientes			
Tipo de prueba	**Nivel mínimo de escala de medición de la variable**	**Prueba**	**Secciones del libro en que se estudia**
Igualdad de más de 2 medias	Ordinal	Kruskall-Wallis	17.6
Pruebas para más de 2 muestras relacionadas			
Igualdad de más de 2 medias	De intervalo	Friedman para diseños en bloques aleatorizados	17.7

17.2 Prueba de rachas para aleatoriedad de Wald-Wolfowitz

Esta prueba, a la que también se conoce como *prueba de Wald-Wolfowitz* o *prueba de corridas (rachas)* se utiliza para determinar si un conjunto de datos es aleatorio o no, como en el caso de datos obtenidos para muestras aleatorias. Otros ejemplos podrían ser las tablas de números aleatorios que se usan para extraer muestras o los resultados de sorteos. En el caso de muestreo para control de la calidad en producción, se utiliza esta prueba para determinar si las observaciones muestrales de la variable o las variables con las que se mide la calidad tienen un comportamiento aleatorio. Se puede aplicar dicha prueba, como se verá en un ejemplo de esta sección, para determinar si las observaciones de una gráfica de control son o no aleatorias, pues la aparición de patrones sistemáticos o periódicos en su selección es señal de deficiencias que es necesario corregir.

Esta prueba se basa en una variable dada en escala nominal, ya que se clasifican los datos en una de 2 categorías: defectuoso, no defectuoso; hombre, mujer; menor a la mediana, superior o igual a la mediana, etc. Es más, como se trata de una variable nominal que sólo tiene 2 categorías se trata en particular de una variable binomial.

Racha (o corrida). Es una sucesión ininterrumpida de casos del mismo tipo.

Esta prueba se basa en el número de rachas o corridas que se aprecian en la muestra; se puede definir una **racha** o **corrida** como una sucesión ininterrumpida de casos del mismo tipo.

Por ejemplo, si la variable que se mide es el sexo, y se utiliza "M" para representar masculino y "F" para femenino, se podría tener una muestra con las siguientes observaciones:

MMMMMFFFMMMFFFFMFMMMMFFFFFF

Entonces, la primera racha sería la sucesión inicial de 5 elementos del sexo masculino, esta primera racha se cortaría por la primera aparición de un caso del sexo femenino; la segunda racha empieza con la primera aparición de un caso femenino y constaría de 3 elementos, la racha siguiente se formaría por los 3 casos de M que aparecen después, y así sucesivamente. El razonamiento básico en que se apoya esta prueba es que si las observaciones son verdaderamente aleatorias, es prácticamente imposible que haya sólo 2 rachas:

MMMMMMMMMMMMMMMFFFFFFFFFFFFFFFFFFF

o que haya una alternancia perfecta de los 2 valores de la variable:

MFMFMFMFMFMFMFMFMFMFMFMFMFMFMFMF

Como se verá más adelante, lo que se hace es, primero, identificar las rachas (para lo cual conviene subrayarlas) para después contarlas. Cabe mencionar que se acostumbra representar al número de rachas con una r. Por otro lado, como las observaciones de la muestra que tiene n elementos se dividen en las 2 categorías de la variable, se hablaría de n_1 y de n_2, que serían el número de elementos en cada categoría. En el ejemplo anterior, n_1 sería el número de elementos del sexo masculino y n_2 el número de elementos del sexo femenino y se debe cumplir, por supuesto, que

$$n = n_1 + n_2$$

17.2.1 Características

Las principales características de esta prueba son:

1. Hipótesis nula: como es una prueba sobre aleatoriedad, no se hacen inferencias sobre parámetros específicos.
2. Nivel de medición: puede ser desde el nivel de escala nominal.
3. Suposiciones: ninguna.
4. Extremos: se pueden realizar pruebas de 1 o 2 extremos.
5. Número de muestras: se aplica sólo a una muestra.
6. Estadístico de prueba. Tradicionalmente se dividía la aplicación de esta muestra en 2: para muestras pequeñas, cuando, $n_1 \leq 10$, o $n_2 \leq 10$, se utilizaban valores críticos ofrecidos en tablas adaptadas de otras publicadas en 1943.

Para muestras relativamente grandes, cuando $n_1 > 10$ y $n_2 > 10$, se sabe que la r se distribuye en forma aproximadamente normal, con:

$$_r = \frac{2n_1\ n_2}{n_1 + n_2} + 1 \qquad \textbf{(17.1)}$$

y

$$\sigma_r = \sqrt{\frac{2n_1\ n_2\left(2n_1\ n_2 - n_1 - n_2\right)}{\left(n_1 + n_2\right)^2\left(n_1 + n_2 - 1\right)}} \qquad \textbf{(17.2)}$$

Por lo que se puede utilizar la z normal como estadístico de prueba:

$$Z = \frac{r - \mu_r}{\sigma_r} \qquad \textbf{(17.3)}$$

Esta aproximación es válida porque la distribución muestral de r tiende a ser normal conforme aumenta el tamaño de la muestra. Se considera que la aproximación es aceptable cuando tanto n_1 como n_2 son mayores a 10; o sea, esta prueba sólo se puede aplicar cuando ambas n son mayores de 10.

Sin embargo, se conoce la manera de calcular la probabilidad de ocurrencia de cualquier número de rachas, independientemente de los tamaños de las 2 partes en las que se divide la muestra: n_1 y n_2, y que con la ayuda de Excel es considerablemente fácil de calcular.

Este procedimiento en Excel tiene la ventaja de que es el mismo sin importar el tamaño de la muestra. En la forma tradicional de abordar este tipo de pruebas es necesario entender y aplicar 2 procedimientos: uno para muestras grandes y otro para muestras pequeñas.

Otra ventaja adicional es que, en el método tradicional, se podían utilizar solamente los niveles de significación que las tablas disponibles contemplaban y que, por lo general, era uno solo, 1 o 5%. Con la utilización del procedimiento que se propone aquí, basado en probabilidades, se puede emplear cualquier nivel de significación que se desee.

El procedimiento consiste en calcular la probabilidad de obtener la cantidad de rachas que se observa en la muestra y que se calcula con las combinaciones vistas en el tema de probabilidad, en la sección 4.3. Se calculan probabilidades para números pares y nones de rachas, de la siguiente manera:

Para número par de rachas, cuando $r = 2k$, en donde k es un número entero positivo:

$$P(r) = \frac{2\binom{n_1 - 1}{k-1}\binom{n_2 - 1}{k-1}}{\binom{n_1 + n_2}{n_1}} \tag{17.4}$$

Para un número non de rachas, con $r = 2k + 1$:

$$P(r) = \frac{\binom{n_1 - 1}{k}\binom{n_2 - 1}{k-1} + \binom{n_1 - 1}{k-1}\binom{n_2 - 1}{k}}{\binom{n_1 + n_2}{n_1}} \tag{17.5}$$

En ambas fórmulas los paréntesis indican combinaciones. Se presentan ahora unos ejemplos.

Ejemplo 17.1

Las preguntas de un examen sólo aceptan respuestas verdaderas (V) o falsas (F). El examen consta de 13 preguntas y la sucesión ordenada de respuestas correctas es:

FFVVVFFVVFFVF

con un nivel de significación de 0.01, ¿se puede considerar que la sucesión de respuestas verdaderas y falsas es aleatoria?

Solución:

1. Las hipótesis:
   H_0: La sucesión de F y V es aleatoria.
   H_1: La sucesión no es aleatoria.

2. El valor muestral del estadístico de prueba, el número de rachas, r: se identifican, se cuentan las rachas y se determina n_1, el número de respuestas verdaderas y n_2, el número de respuestas falsas.

FF <u>VVV</u> FF <u>VV</u> FF <u>V</u> F

$$r = 7$$
$$n_1 = 6$$
$$n_2 = 7$$

Al calcular ahora la probabilidad de obtener este número de rachas, con esos tamaños de submuestras, y muestra de 13 observaciones:

Como las rachas son 7, y es número non, sustituyendo en:

$$r = 2k + 1$$
$$7 = 2k + 1$$
$$2k = 7 - 1$$
$$k = 3$$

Sustituyendo estos valores en la fórmula (17.5).

$$P(r)=\frac{\binom{n_1-1}{k}\binom{n_2-1}{k-1}+\binom{n_1-1}{k-1}\binom{n_2-1}{k}}{\binom{n_1+n_2}{n_1}}$$

$$=\frac{\binom{6-1}{3}\binom{7-1}{3-1}+\binom{6-1}{3-1}\binom{7-1}{3}}{\binom{6+7}{6}}$$

$$=\frac{\binom{5}{3}\binom{6}{2}+\binom{5}{2}\binom{6}{3}}{\binom{13}{6}}$$

Recuerda que la forma de resolver estas combinaciones (para un repaso, ir a la sección 4.3) es:

$$\frac{\binom{5}{3}\binom{6}{2}+\binom{5}{2}\binom{6}{3}}{\binom{13}{6}}$$

$$=\frac{\frac{5\cdot4\cdot3}{3\cdot2}\cdot\frac{6\cdot5}{2}+\frac{5\cdot4}{2}\cdot\frac{6\cdot5\cdot4}{3\cdot2}}{\frac{13\cdot12\cdot11\cdot10\cdot9\cdot8}{6\cdot5\cdot4\cdot3\cdot2}}=0.20396$$

Aunque laboriosas, las operaciones son muy sencillas, y mucho más si se resuelven con Excel, tal como se apreciará en la sección siguiente.

17.2.2 Excel y el cálculo de probabilidades para la prueba de rachas de Wald-Wolfowitz

	A	B	C
1		VALORES	COMBINACIONES
2	n1	6	=COMBINAT(B2-1,B4)
3	n2	7	=COMBINAT(B3-1,B4-1)
4	k	3	=COMBINAT(B2-1,B4-1)
5			=COMBINAT(B3-1,B4)
6			=COMBINAT(B2+B3,B2)
7			
8	PROBABILIDAD (pares)	=2*C4*C3/C6	
9			
10	PROBABILIDAD (nones)	=(C2*C3+C4*C5)/C6	
11			

Figura 17.1 Hoja de Excel con funciones para calcular probabilidades en aplicaciones de la prueba de rachas para aleatoriedad.

	A	B	C
1		VALORES	COMBINACIONES
2	n1	6	10
3	n2	7	15
4	k	3	10
5			20
6			1716
7			
8	PROBABILIDAD (pares)	0.17482517	
9			
10	PROBABILIDAD (nones)	0.2039627	
11			

Figura 17.2 Hoja de Excel con datos para calcular probabilidades en aplicaciones de la prueba de rachas para aleatoriedad.

Si se arma una hoja de cálculo de Excel como la que aparece en la figura 17.1, es sumamente sencillo hacer los cálculos.

En la figura 17.1 se anotaron los valores de las variables en las celdas B2 a B4 y en las celdas C2 a C6 se incorporaron las funciones de las combinaciones que se requieren para hacer los cálculos para los 2 casos de esta prueba de rachas, para rachas impares y rachas pares.

Ahora, en la figura 17.2 se muestra la misma tabla de Excel pero ya no mostrando las fórmulas con la función "COMBINAT" de Excel, sino con sus resultados. Si se comparan las fórmulas que ahí aparecen se verá que coinciden con las combinaciones que se requieren en las fórmulas (17.1) y (17.2) para calcular las probabilidades, dichas fórmulas se anotan en las celdas B8 y B10.

Como se aprecia, dado que ambas fórmulas comparten cálculos de combinaciones, una vez que se anotan los datos aparecen los 2 resultados, sólo que en el caso del ejemplo 17.1 se sabe que el valor buscado es de rachas nones, 0.2039627, porque así se determinó.

Finalmente, como la probabilidad de obtener 7 rachas en las condiciones planteadas es de 20.7%, que es muy superior al nivel de confianza especificado, no se rechaza la hipótesis nula y se concluye que la sucesión de respuestas verdaderas y falsas en ese examen es aleatoria.

Tal como se había anotado antes, este procedimiento es mucho más sencillo que analizar 2 técnicas distintas para muestras grandes y pequeñas y, por ello, se sugiere enfáticamente su utilización.

Para quien desee consultar esos otros 2 procedimientos, se sugiere revisar cualquier texto de estadística que incluya el tema de estadística no paramétrica y, para una revisión de la forma en la que se derivan las fórmulas (17.1) y (17.2), tema que rebasa el alcance de este texto, se sugiere consultar el libro *Estadística matemática*, de John E. Freund y Ronald E. Walpole, editado por Prentice Hall, México, 1990.

EJEMPLO 17.2

Se determinó el número de artículos defectuosos por hora durante las horas en las que se trabaja en una fábrica. Los resultados se presentan en la tabla 17.2.

Tabla 17.2 Artículos defectuosos por hora en un proceso

Hora	Núm. de artículos defectuosos *X*	Posición de *X* respecto a la Med = 3.5
6-7 a.m.	6	S
7-8 a.m.	3	I
8-9 a.m.	1	I
9-10 a.m.	5	S
10-11 a.m.	0	I
11-12 a.m.	8	S
12-1 p.m.	4	S
1-2 p.m.	1	I
2-3 p.m.	7	S
3-4 p.m.	1	I
4-5 p.m.	9	S
5-6 p.m.	8	S
6-7 p.m.	2	I
7-8 p.m.	9	S
8-9 p.m.	1	I
9-10 p.m.	0	I
10-11 p.m.	6	S
11-12 p.m.	3	I

¿Se puede concluir, con un nivel de significación del 0.05, que la aparición de artículos defectuosos se da al azar en todas las horas?

Solución:

1. H_0: Los productos defectuosos se dan al azar, cada hora, en el proceso.
   H_1: Los productos defectuosos no aparecen al azar.
2. El valor muestral de la medida.

Para poder determinar si los productos defectuosos aparecen al azar es necesario determinar, en primer lugar, los casos que están por encima y por debajo de la mediana, entonces se empieza por determinarla. Para hacerlo, en seguida se ordenan los datos de menor a mayor

Número de artículos defectuosos, de menor a mayor

0, 0, 1, 1, 1, 1, 2, 3, 3, 4, 5, 6, 6, 7, 8, 8, 9, 9.

como $n = 18$, la mediana es el valor que ocupa el lugar 9.5

$$\text{Posición de la mediana} = \frac{n+1}{2} = \frac{18+1}{2} = 9.5$$

es entonces el valor encontrado entre 3 (el noveno elemento) y 4 (el décimo elemento), por lo que la mediana es:

$$\text{Med} = \frac{3+4}{2} = 3.5$$

En la tercera columna de la tabla 17.2 se identifica si el número de artículos defectuosos es superior (S) o inferior (I) a la mediana. Se reproduce en seguida la sucesión de varios artículos (S e I) y se identifican las rachas siguientes

$$\underline{S}\ I\ I\ \underline{S}\ I\ \underline{S\ S}\ I\ \underline{S}\ I\ \underline{SS}\ I\ \underline{S}\ I\ I\ \underline{S}\ I$$

$r = 14$.
$n_1 = 9$, número de casos de artículos superiores a la mediana.
$n_2 = 9$, número de casos de artículos inferiores a la mediana.

Como $r = 14$ es par, se utiliza la fórmula 17.1 para calcular la probabilidad de obtener 14 rachas, con los valores anotados antes y con $k = 7$. Utilizando la misma hoja de Excel que se ilustró en las figuras 17.1 y 17.2 se obtiene una probabilidad de 0.0322 y, como esta probabilidad es inferior al nivel de significación de 0.05, se rechaza la hipótesis nula y se concluye que la aparición de artículos defectuosos en las diferentes horas en las que se trabaja en esa fábrica no es aleatoria.

EJERCICIOS 17.2 Pruebas de rachas para aleatoriedad

1. Un observador apostado en el límite entre el Distrito Federal y el Estado de México anotó, para una serie de 60 automóviles que pasaron frente a él durante cierto periodo, si los automóviles tenían placas del Distrito Federal o de cualquier otra entidad federativa y obtuvo los siguientes resultados (D = Distrito Federal y F = foráneo):

DDFDDDDFFDDDDFDFFDDDDFDFFDDDDDFDDDFDF
DDDDFFDFFFFDDDDFDFFDDDF

A un nivel de significación de 5%, ¿se puede afirmar que durante ese periodo los automóviles con placas del DF y los que tenían placas de otras entidades pasaron por ese punto de una manera verdaderamente aleatoria?

2. Se observó que los pesos, en gramos, de 30 latas de duraznos en almíbar que fueron empacados por una máquina fueron: 493, 487, 444, 419, 494, 436, 417, 481, 409, 450, 433, 488, 415, 454, 402, 453, 403, 455, 457, 451, 444, 449, 459, 460, 445. 452, 465, 470, 466 y 453. En principio, se trata de latas de 450 gramos. Si se asigna la letra D (deficiente) a las latas que pesan menos de 450 gramos y S (suficiente) a las que pesan 450 gramos o más, ¿se puede afirmar, a un nivel de significación de $\alpha = 0.01$, que se trata de una muestra aleatoria?

3. Lance un dado 50 veces y determine la sucesión de números que aparecen. Representando con G los números

grandes (4, 5 y 6) y con C los números chicos (1, 2 y 3), ¿se puede considerar que se trata de una muestra aleatoria? Utilice un nivel de significación de 0.05.

4. Lance una moneda 40 veces y registre los resultados. ¿Se puede considerar que son aleatorios a un nivel de significación de 0.01?

5. Registre los resultados del premio mayor de la Lotería Nacional de los últimos 40 sorteos. Identifique como P (par) y como I (impar) al último dígito de cada número ganador. ¿Las terminaciones de estos primeros premios aparecieron al azar? Utilice un nivel de significación de 0.01.

6. Se registraron los resultados de los colores (Rojo o Negro) que aparecieron en un periodo determinado en una mesa de ruleta de un casino, y se obtuvieron los siguientes resultados: R,R,R,N,N,N,R,N,R,R,R,R,N,N,R,N,N,N, R,R,R,R,N,N,N,N,N,R,R,R,N,N,R,N,R,N,R,R,R,R.¿Se puede considerar que los resultados aparecieron al azar? Utilice $\alpha = 0.05$.

17.3 Prueba de los signos

Esta prueba implica una hipótesis nula sobre la mediana. Como se puede aplicar a una sola muestra o a 2 muestras apareadas (relacionadas), la hipótesis nula entonces puede referirse a 1 o 2 medianas poblacionales. Esta prueba equivale a una prueba paramétrica sobre la media de una población y, al igual que en ésta, se pueden realizar pruebas de 1 o de 2 extremos.

17.3.1 Características

Las principales características de esta prueba son:

1. Hipótesis nula: la prueba se hace sobre la o las medianas poblaciones aunque el procedimiento implica probar una hipótesis equivalente sobre proporciones poblacionales.
2. Nivel de medición: la variable debe estar cuando menos en escala ordinal.
3. Suposiciones: ninguna.

Los signos a los que hace referencia el nombre de esta prueba son el de adición (+) y de sustracción (–) que se utilizan para determinar si los valores observados están encima (+) o debajo (–) de la mediana para el caso de una muestra. En el caso de 2 muestras se adjudican los signos según si la observación de la primera muestra fuera superior a la observación correspondiente de la segunda muestra (+) o viceversa (–).

Como la mediana es el valor que divide a un conjunto de datos de manera que la mitad de las observaciones se encuentran por debajo de ella y la otra mitad por encima, la prueba se realiza utilizando la distribución binomial, considerando que $\pi = 0.5$.

4. Extremos: se pueden hacer pruebas de 1 o de 2 extremos.
5. Número de muestras: se puede aplicar a una sola muestra o a 2 muestras relacionadas (apareadas).
6. Estadístico de prueba: antes de explicar las diferencias entre los procedimientos para muestras grandes y muestras pequeñas, se desea resaltar aquí que, con el criterio de que una muestra ya es grande cuando tiene cuando menos 10 elementos, en realidad, el procedimiento para pruebas pequeñas, que es más laborioso que el de muestras grandes, en la práctica se aplica a muestras muy pequeñas.

   a) Para muestras pequeñas ($n < 10$) se utiliza la distribución binomial para determinar los valores críticos, en otras palabras, la decisión se toma en términos de la distribución de probabilidad binomial.

   b) Para muestras grandes, cuando la muestra tiene cuando menos 10 elementos, se utiliza la aproximación normal, con z. Ver la sección 9.8, prueba de hipótesis sobre una proporción poblacional:

$$z = \frac{p - \pi}{\sigma_p} = \frac{p - \pi}{\sqrt{\frac{\pi Q}{n}}}$$

O, con valores muestrales:

$$z = \frac{p - \pi}{s_p} = \frac{p - \pi}{\sqrt{\frac{pq}{n}}}$$

Como el valor esperado de signos positivos (y negativos) es 0.5, el error estándar de la proporción se convierte en:

$$s_p = \sqrt{\frac{pq}{n}} = \sqrt{\frac{0.5(0.5)}{n}} = \sqrt{\frac{0.25}{n}}$$

Y la del estadístico de prueba:

$$z = \frac{p - \pi}{s_p} = \frac{p - \pi}{\sqrt{\frac{0.5\,(0.5)}{n}}} = \frac{p - \pi}{\sqrt{\frac{(0.25)}{n}}}$$

Adicionalmente, como los datos sobre los que se aplica la prueba son los signos, y se trata de una prueba binominal, es necesario aplicar una corrección por discontinuidad, que ya se estudió en la sección 6.3 "Ajuste cuando se utiliza la distribución normal para evaluar probabilidades de una variable discreta (ajuste por discontinuidad)", por lo que el estadístico de prueba se convierte en:

$$z = \frac{(p \pm 0.005) - \pi}{\sqrt{\frac{(0.25)}{n}}} \tag{17.6}$$

Es importante observar aquí que el factor de corrección por continuidad es 0.005 y no 0.5 como se manejó antes en la sección 6.3 ya que, en este caso, se usan proporciones que se manejan en decimales, en tanto que en dicha sección se manejaron casos en unidades y la expresión decimal de 0.5 es, precisamente, 0.005. En los ejemplos que siguen se revisan los detalles.

17.3.2 Prueba del signo para una muestra pequeña

Se ilustra este procedimiento con el ejemplo siguiente:

Ejemplo 17.3

Se fabrican artículos mediante un proceso tradicional que logra una producción mediana de 80 artículos por turno. Los trabajadores propusieron un nuevo sistema de "trabajo por equipos" que afirman que al mismo tiempo que les da mayor flexibilidad en el trabajo, permite aumentar el nivel de la producción. En la tabla 17.3 se presentan los resultados de la producción de 15 turnos muestreados al azar. Pruebe con un nivel de significación de 0.05 la afirmación de los trabajadores.

Tabla 17.3 Datos para el ejemplo 17.3

Turno	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Artículos fabricados	77	85	90	78	92	92	89	78	89	80	94	81	76	85	80

Solución:

1. Las hipótesis:
   H_0: $Med \leq 80$.
   H_1: $Med > 80$.
   Aquí hay que observar que esas hipótesis se pueden plantear de manera equivalente como:
   H_0: $\pi \leq 0.5$.
   H_1: $\pi > 0.5$.

   En donde π representa la proporción de turnos en los que se fabrican más de 80 artículos.

2. El valor muestral de la mediana:

Como la prueba se hace en términos de proporciones, en primer lugar se determina si las observaciones (artículos fabricados por turno) es superior (+) o inferior (–) a la mediana supuesta en la hipótesis nula (80) y se deben eliminar los casos en los que ambos valores sean iguales, con la consiguiente reducción del tamaño de la muestra. En la tabla 17.4 se reproducen los datos presentados antes, incluyendo el signo correspondiente a cada diferencia.

Tabla 17.4 Datos del ejemplo 17.3

Turno	Artículos fabricados	Signo de la diferencia (x – 80)
1	77	–
2	85	+
3	90	+
4	78	–
5	92	+
6	92	+
7	89	+
8	78	–
9	89	+
10	80	0
11	94	+
12	81	+
13	76	–
14	85	+
15	80	0

Como se puede ver en la tabla 17.4, hay 2 casos (turnos 10 y 15) en los que el valor de la diferencia es 0, por lo cual se les elimina de toda consideración posterior y, por consiguiente, la muestra se reduce de $n = 15$ a $n = 13$.

Como, por otro lado, son 9 los signos de más, la proporción de éstos es de:

$$p = \frac{x}{n} = \frac{9}{13} = 0.6923077$$

3. El valor crítico del estadístico de prueba:

Para $\alpha = 0.05$, el valor crítico para una prueba de un extremo es precisamente este valor del nivel de significación, ya que para tomar la decisión se va a comparar la probabilidad de obtener los resultados muestrales, el número de turnos que superan a la producción mediana o más, con este nivel de significación y se rechaza la hipótesis nula si dicha probabilidad es inferior al nivel de significación. Por supuesto, si se tratara de una prueba de 2 extremos (que no es el caso aquí) se tendría que dividir α entre 2.

4. El valor esperado de la variable.
5. El error estándar de la variable: no es necesario calcularlos ya que la prueba se hace comparando directamente la probabilidad del nivel de significación con la probabilidad de obtener el número de resultados observados o más.
6. El valor calculado del estadístico de prueba: se determina utilizando la distribución binomial, pues se tiene un caso de muestra relativamente pequeña ($n = 13$). Los valores de la $P(x)$ para diferentes valores de x se pueden determinar utilizando la función de distribución de probabilidad binomial, con las tablas de probabilidades binomiales (tabla 1 del apéndice) o mediante la función "distr.binom" que se vieron en el capítulo 5.

Ahora para determinar el valor calculado del estadístico de prueba, o sea la probabilidad de tener 9 o más turnos con producción superior a la mediana, se deben determinar las probabilidades correspondientes a diferentes valores de x. Los datos para el ejemplo, con $\pi = 0.5$ y $n = 13$ son los que se presentan en la tabla 17.5.

Tabla 17.5 Distribución de probabilidad binomial para el ejemplo 17.3

x	$P(x) = \frac{n!}{x!(n-x)!} p^x q^{n-x}$	$P(x)$ acumulada
0	0.0001	0.0001
1	0.0016	0.0017
2	0.0095	0.0112
3	0.0349	0.0461
4	0.0873	0.1334
5	0.1571	0.2905
6	0.2095	0.5
7	0.2095	0.7095
8	0.1571	0.8666
9	0.0873	0.9539
10	0.0349	0.9888
11	0.0095	0.9983
12	0.0016	0.9999
13	0.0001	1.0

El significado de los valores de la tabla es el siguiente: 0.1571 del renglón correspondiente a $x = 8$ significa que, dadas las condiciones ($\pi = 0.5$ y $n = 13$), existe una probabilidad de 0.1571 (o de 15.71%) de que una muestra al azar de 13 turnos arroje un resultado de exactamente 8 turnos con menos (o más, que es lo mismo) de 80 artículos fabricados. De acuerdo con esta interpretación, la probabilidad acumulada del renglón $x = 8$ significa que si se elige una sola muestra al azar, se tiene una probabilidad del 0.8666 (u 86.66%) de que el número de turnos en los que se fabrican menos de 80 artículos sea de 8 o menos. Al mismo tiempo, como la suma de las probabilidades de todos los resultados posibles debe ser igual a 1, se puede decir que la probabilidad de que haya más de 8 (o, en otras palabras, 9 o más) turnos con producción superior a la mediana es de $1 - 0.8666 = 0.1344$, o 13.44 por ciento.

Ahora, como lo que se trata de probar en la hipótesis alternativa es que el número mediano de artículos fabricados es superior a 80 artículos por turno, se rechaza la hipótesis nula si la probabilidad del número de turnos o más que superan a la mediana hipotética es menor que el correspondiente nivel de significación, o sea, para poder rechazar la hipótesis nula de que el número mediano de artículos fabricados por turno sigue siendo de 80 o menos, el valor resultante de la muestra debe ser lo suficientemente grande para mostrar que no concuerda con esa suposición.

7. La decisión: como se observaron en la muestra 9 turnos que superan a la producción mediana, se puede apreciar en la tabla 17.5 que la probabilidad de tener estos 9 turnos, o más, es de

$$P(x \geq 9) = 0.0873 + 0.0349 + 0.0095 + 0.0016 + 0.0001 = 0.1334.$$

Este valor se puede calcular también a partir de las probabilidades acumuladas:

$$P(x \geq 9) = 1 - 0.8666 = 0.1334.$$

Como esta probabilidad es superior a 0.05 del nivel de significación, se tiene que el valor muestral cae en la región de aceptación, no se rechaza la hipótesis nula y se concluye, por lo tanto, que el nuevo proceso de trabajo en equipos no aumentó la mediana del número de artículos fabricados por turno.

17.3.3 Prueba del signo para una muestra grande (aproximación normal)

En el ejemplo siguiente se ilustra el procedimiento para realizar esta prueba de hipótesis para una muestra grande, mediante una aproximación con la distribución normal.

Ejemplo 17.4

Aunque en el ejemplo 17.3 se utilizó la distribución binomial para hacer la prueba, considerando que se tenía una muestra relativamente pequeña con 13 elementos, después de eliminar los 2 casos en los que la mediana resultó ser igual a 80, como el criterio para utilizar la aproximación normal es que la muestra sea de cuando menos 10 elementos, y en este caso sí se cumple, se resuelve ahora el mismo ejemplo utilizando la distribución normal.

Solución: Las hipótesis son las mismas:

$H_0: Med \leq 80$
$H_1: Med > 80$

o:

$H_0: \pi \leq 0.5$
$H_1: \pi > 0.5$

Se utiliza también el mismo nivel de significación de 0.05.

Se redujo la muestra a $n = 13$, al eliminar 2 casos en los que la mediana resultó ser igual a 80. La proporción de signos + fue:

$$p = \frac{x}{n} = \frac{9}{13} = 0.6923077$$

El valor crítico del estadístico de prueba, z, para una prueba de un solo extremo es 1.645, ya que $P(z \geq 1.645) = 0.05$.

El valor calculado o empírico del estadístico de prueba:

$$z = \frac{(p \pm 0.005) - \pi}{\sqrt{\frac{(0.25)}{n}}} = \frac{(0.6923 - 0.005) - 0.5}{\sqrt{\frac{0.25}{13}}}$$

$$= \frac{0.6873 - 0.5}{0.1387} = 1.35$$

Como este valor empírico de z es menor que el valor crítico de 1.645, no es posible rechazar la hipótesis nula y, al igual que en el ejemplo 17.3, se concluye que el nuevo proceso de trabajo en equipos no aumentó la mediana del número de artículos fabricados por turno.

NOTA

En este caso se restó el factor de corrección por discontinuidad debido a que, como se quería probar alternativamente que la mediana era superior a 80, los valores eran suficientemente altos y actuarían en favor de la hipótesis nula; por ello, cualquier unidad que se cuente hacia la parte superior de la curva normal empieza en su lado izquierdo. Por eso se hizo la resta, para incluir la parte izquierda.

17.3.4 Prueba del signo para 2 muestras apareadas pequeñas

El concepto de 2 muestras apareadas se refiere a que se toman 2 mediciones diferentes a los mismos elementos, como en el caso del ejemplo siguiente en el que se obtuvieron de los mismos 20 alumnos 2 calificaciones para 2 profesores diferentes. Otra versión de este mismo caso se presentaría al pedir a 2 grupos de personas que den una evaluación sobre el mismo conjunto de personas (objetos, métodos, etcétera).

Esta prueba se realiza de la misma manera que para el caso de una sola muestra, con la única diferencia de que, con 2 muestras, los signos se obtienen como la diferencia entre el valor de la primera muestra y el correspondiente valor de la segunda. También, al igual que se vio antes, si hay empates de observaciones (diferencias de 0) se eliminan los correspondientes casos.

Ejemplo 17.5

Se le pidió a un grupo de 20 alumnos que calificaran el desempeño de 2 profesores, de acuerdo con varios criterios, y en una escala de 1 a 10. Se obtuvieron los resultados que se muestran en la tabla 17.6.

Tabla 17.6 Datos y cálculos para el ejemplo 17.5

	Calificación		
Alumno	Profesor A x_1	Profesor B x_2	Signo de $(x_1 - x_2)$
1	7	9	–
2	5	6	–
3	8	5	+
4	9	8	+
5	3	4	–
6	8	5	+
7	10	10	0
8	8	9	–
9	3	6	–

(*continúa*)

Tabla 17.6 (*continuación*)

	Calificación		
Alumno	Profesor A x_1	Profesor B x_2	Signo de $(x_1 - x_2)$
10	5	4	+
11	7	10	–
12	9	6	+
13	5	3	+
14	4	4	0
15	7	9	–
16	10	10	0
17	10	9	+
18	5	8	–
19	5	4	+
20	6	6	0

Probar, a un nivel de significación de 0.05 la hipótesis nula de que no existe diferencia entre las calificaciones asignadas por los alumnos a los 2 profesores.

Solución:

1. Las hipótesis:

H_0: $Med_1 = Med_2$
H_1: $Med_1 \neq Med_2$

o, de manera equivalente

H_0: $\pi_1 = \mu_2$
H_1: $\pi_1 \neq \mu_2$

En donde π representa las proporciones de estudiantes que dieron una mayor calificación al profesor A, aunque, al igual que en todos los casos de variables binomiales, se podría asumir la posición desde el otro punto de vista: la proporción de estudiantes que asignaron una mayor calificación al profesor B ya que, a final de cuentas, el resultado de la prueba es el mismo.

2. El valor crítico del estadístico de prueba:

Para $\alpha = 0.05$, el valor crítico para una prueba de 2 extremos se determina utilizando la distribución binomial con $n_1 = n_2 = 16$, debido a que, de los 20 elementos (alumnos) originales, se eliminan los 4 casos en los que coinciden las calificaciones para los 2 profesores.

Los valores de la $P(x)$ para diferentes valores de x se pueden determinar utilizando la función de distribución de probabilidad binomial o mediante las tablas correspondientes o mediante la función "Distr.Binom" de Excel, de la misma manera que se vio en la sección para una sola muestra.

En la misma tabla 17.6 se anotaron los signos correspondientes a cada par de observaciones y, como se trata de una prueba de 2 extremos, se divide el nivel de significación entre 2:

$$\frac{\alpha}{2} = \frac{0.05}{2} = 0.025$$

Volviendo a la tabla 17.6 se observa que hubo 10 signos +, por lo que es necesario calcular las probabilidades binomiales de 10 o más éxitos (signos +) para comparar este valor con el nivel de significación. En la tabla 17.7 se muestran los valores de las probabilidades binomiales con $P = 0.5$ y $n = 16$ y para valores de x de 10 a 16, con su correspondiente suma, en donde se puede ver que la probabilidad de obtenerse 10 o más signos positivos es 0.2272.

Tabla 17.7 Probabilidades binomiales con $P = 0.5$ y $n = 16$

10	0.1221924
11	0.0666504
12	0.0277710
13	0.0085449
14	0.0018311
15	0.0002441
16	0.0000153
Suma	0.2272491

Como el valor de la probabilidad de obtener 10 signos positivos, o más, es de 0.22, y este valor es mayor que el nivel de significación estipulado, 0.025, no se rechaza la hipótesis nula y se concluye que la opinión de los alumnos es la misma para los 2 profesores.

17.3.5 Prueba del signo para 2 muestras apareadas grandes (aproximación normal)

Igual que se hizo con el caso de una sola muestra, también aquí el ejemplo anterior para muestras pequeñas se resolvió con la distribución binomial, a pesar de que se tenía una muestra con más de 10 elementos, circunstancia que permite el uso de la aproximación normal.

Se hizo esto con 2 propósitos: en primer lugar, para mostrar que con ambas técnicas se llegan a las mismas conclusiones y, en segundo término, para enfatizar el hecho de que el criterio de que una muestra pequeña es de 9 elementos o menos implica, en realidad, no sólo una muestra pequeña sino, más bien, muy pequeña.

Se resuelve a continuación el ejemplo anterior, ahora como ejemplo 17.6, utilizando la aproximación normal.

Ejemplo 17.6

El ejemplo trata de un grupo de 20 alumnos que calificó a 2 profesores y se desea probar que no existe diferencia entre las calificaciones asignadas por los alumnos a los 2 profesores con un nivel de significación de $\alpha = 0.05$.

Solución:

Las hipótesis:

H_0: $Med_1 = Med_2$
H_1: $Med_1 \neq Med_2$

o, de manera equivalente,

H_0: $\pi_1 = \mu_2$
H_1: $\pi_1 \neq \mu_2$

Se encontró que, tras eliminar 4 casos cuyas calificaciones eran iguales para los 2 profesores, restó una muestra de 16 alumnos, de los cuales 10 arrojaron signos positivos en sus evaluaciones de los maestros. Por ello, la proporción de signos + fue:

$$p = \frac{x}{n} = \frac{10}{16} = 0.625$$

El valor crítico del estadístico de prueba, z, para una prueba de 2 extremos es 1.96, ya que $P(-1.96 \leq z \leq 1.96) = 0.05$.

El valor calculado o empírico del estadístico de prueba:

$$z = \frac{(p \pm 0.005) - \pi}{\sqrt{\frac{0.25}{n}}} = \frac{(0.625 - 0.005) - 0.5}{\sqrt{\frac{0.25}{16}}} = \frac{0.62 - 0.5}{0.125} = 0.96$$

Y, como este valor empírico de z es menor que el valor crítico de 1.96, no es posible rechazar la hipótesis nula y, al igual que en el ejemplo 17.5, se concluye que la opinión de los alumnos es la misma para los 2 profesores.

NOTA

En este caso se restó el factor de corrección por discontinuidad, debido a que, tratándose de una prueba de 2 extremos, y observando que la proporción de signos positivos era mayor de 0.5, en su caso, el rechazo de la hipótesis nula se daría con valores altos de la z empírica y, por ello, cualquier unidad que se cuente hacia la parte superior de la curva normal, empieza en su lado izquierdo.

Ejercicios 17.3 Prueba de los signos

Prueba de los signos para una muestra pequeña

1. Un gerente de personal desea evaluar si aumentó el nivel de satisfacción en el trabajo, desde una mediana histórica de 70. Toma una muestra aleatoria de 8 empleados y obtiene las siguientes calificaciones: 95, 68, 86, 87, 99, 64, 68 y 78. Con un nivel de significación de 0.01, ¿los datos que obtuvo confirman su hipótesis?
2. Durante los 5 meses de la temporada alta de turismo, un parque de diversiones registró 15, 13.5, 16.50 19.5 y 10.5 miles de boletos vendidos. Si la mediana para toda la red nacional de parques de diversiones es de 19 mil boletos por mes, ¿se puede afirmar que la afluencia de visitantes a este parque en particular está por debajo de la mediana, con un nivel de significación de 0.01?
3. Los siguientes son los salarios medianos por sector en México en 2011. ¿Se puede decir, con un nivel de significación del 0.01 que el salario mediano general de estos profesionales (que no de la población en general) supera los $6 000 que un funcionario mexicano afirma son suficientes para vivir holgadamente?

Extractiva	$7 588.00
Servicios profesionales	$6 781.00
Educación y salud	$6 378.00
Gobierno	$6 037.00
Transporte	$5 532.00
Construcción	$4 506.00
Transformación	$4 085.00
Comercio	$3 895.00
Turismo	$3 787.00
Servicios personales	$3 164.00
Agricultura	$2 146.00

4. Se rediseña una línea de producción con el propósito de incrementar la producción, cuya mediana actual es de 80 unidades ensambladas por turno. Los resultados de una muestra aleatoria de 8 turnos, tomada después del rediseño arroja las siguientes cantidades de unidades ensambladas: 75, 85, 80, 90, 91, 88, 96, 76. ¿Puede el ingeniero que hizo el rediseño afirmar que su proyecto tuvo éxito con un nivel de significación de 5 por ciento?
5. Se desea evaluar si el contenido de cierta presentación de cereal tiene una mediana de 335 gramos. Para hacerlo, se toma una muestra aleatoria de 9 cajas de cereal y se encuentran los siguientes pesos, en gramos: 334, 331, 343, 337, 340, 331, 340, 334 y 330. Haga la prueba con un nivel de significación de 1 por ciento.

Prueba de los signos para una muestra grande

6. En un puesto carretero de radar de verificación de velocidad de vehículos se encontró que, de 1 000 registros, 975 viajaban a mayor velocidad que la máxima permitida

de 80 kilómetros por hora, 11 viajaban a 80 km/h y 14 a menos de la velocidad máxima permitida. ¿Se puede afirmar, con un nivel de significación de 5%, que esa velocidad máxima permitida es exageradamente reducida?

7. Con el propósito de que le asignen más recursos, un político argumenta que el ingreso mediano por familia en su municipio es inferior a todos los del estado, cuya mediana es de $4 000. Se toma una muestra aleatoria de 15 familias y se obtienen los siguientes resultados: 7 100, 5 600, 4 200, 8 700, 1 000, 3 100, 5 200, 2 000, 15 600, 6 100, 5 000, 2 000, 1 500, 4 100 y 5 900. Con un nivel de significación de 5%, ¿se puede sostener la petición de fondos del político?

8. Una empresa desea comprobar la suposición de que la mediana de sus ingresos semanales por ventas es de cuando menos $12 500. Para hacerlo, toma una muestra aleatoria de 12 semanas y obtiene los siguientes resultados.

Semana	Ventas
1	15 000
2	12 500
3	17 500
4	12 125
5	10 500
6	17 750
7	11 250
8	13 750
9	14 375
10	12 500
11	13 000
12	12 000

Pruebe la suposición de la empresa con un nivel de significación de 0.01.

9. En una clínica médica aseguran que el tiempo de espera de los pacientes antes de la consulta con el especialista no rebasa los 25 minutos. Una muestra de 20 pacientes arrojó los siguientes tiempos de espera: 23, 29, 32, 39, 38, 24, 30, 13, 31, 16, 24, 31, 21, 39, 29, 25, 34, 24, 35 y 20. ¿Se puede sostener la afirmación de la clínica con un nivel de significación de 0.05?

10. Los datos siguientes representan los tiempos (en minutos) que requirió una muestra de 20 técnicos para realizar determinada tarea, después de haber participado en un programa de capacitación para acelerar la realización del trabajo: 18.2, 20.4, 18.3, 15.7, 22.6, 16.9, 17.6, 16.9, 18.2, 17.0, 19.2, 16.7, 19.6, 18.6, 20.2, 18.8, 19.1, 17.6, 18.4 y 18.2. De estudios similares anteriores se sabe que la mediana histórica del tiempo necesario para realizar esta tarea es de 19.5 minutos. Con un nivel de significación de 0.01, ¿se puede afirmar que aumentó la eficiencia de los técnicos con el programa de capacitación?

11. El fabricante del líquido limpiador "El bueno" afirma que el contenido de líquido limpiador en sus botellas de un litro (según la etiqueta) contienen en promedio 1.05 litros. Para verificar esta afirmación, un funcionario del Instituto de Protección al Consumidor elige al azar 15 recipientes de un lote numeroso de envases de un litro de "El bueno" y verifica los contenidos, obteniendo los siguientes resultados en litros: 0.95, 1.0, 1.25, 1.0, 0.9, 1.05, 0.85, 1.0, 1.1, 0.95, 1.15, 0.93, 1.05, 0.97 y 1.03. ¿Se puede concluir, a partir de estos resultados, que la afirmación del fabricante es falsa? Utilice un nivel de significación de 0.05.

Prueba de los signos para 2 muestras pareadas pequeñas

12. Con el propósito de evaluar la efectividad de un curso extracurricular de redacción, se tomó una muestra aleatoria de 9 estudiantes a los que se les aplicaron 2 exámenes, uno antes y otro después del curso de redacción, y se obtuvieron las siguientes calificaciones:

Estudiante	Calificación antes del curso extracurricular	Calificación después del curso extracurricular
1	70	85
2	75	98
3	60	75
4	68	80
5	66	66
6	72	69
7	62	79
8	65	80
9	70	85

Con un nivel de significación de 0.05, ¿hay evidencia para concluir que el curso extracurricular les sirvió a los estudiantes?

13. Los valuadores del Nacional Monte de Piedad tasan las prendas que se presentan para empeño y con base a esta evaluación ofrecen créditos prendarios. Se pidió a 2 valuadores elegidos al azar que tasaran 10 piezas de joyería, obteniéndose los siguientes resultados:

Joya	Valuador A	Valuador B
1	36.4	35.0
2	48.3	46.7
3	40.1	37.2
4	54.8	50.5
5	28.6	29.2
6	42.9	41.2

Joya	Valuador A	Valuador B
7	36.2	35.5
8	38.9	39.1
9	45.2	46.4
10	47.3	46.7

Utiliza la prueba del signo para probar si uno de los 2 valuadores tiende a ser más conservador al tasar las prendas. Realizar la prueba con un nivel de significación de 0.05.

14. En una investigación de mercado se conformó un panel de consumidores para evaluar las preferencias sobre 2 marcas de café instantáneo, y se le pidió a cada miembro del panel asignar una calificación a las 2 marcas, con los siguientes resultados:

Panelista	Marca A	Marca B
1	70	66
2	74	76
3	78	68
4	74	67
5	70	70
6	79	71
7	69	73

Utilizando un nivel de significación de 1%, prueba la hipótesis de que no hay diferencia entre las preferencias entre ambas marcas de café.

15. Se probó una dieta para bajar de peso en 8 personas, con los siguientes resultados:

Persona	Peso antes	Peso después
1	85	78
2	69	63
3	74	71
4	75	73
5	79	80
6	82	80
7	79	77
8	70	70

¿Se puede decir que la dieta es efectiva para bajar de peso, con un nivel de significación de 0.05?

16. En la Facultad de Contaduría y Administración de una universidad se desea evaluar si existe diferencia entre la eficiencia de los estudiantes en los cursos de estadística descriptiva y estadística inferencial. Se eligen al azar 9 estudiantes y se observan los siguientes resultados:

Estudiante	Calificación	
	Estadística descriptiva	Estadística inferencial
1	97	92
2	78	80
3	85	82
4	84	83
5	92	96
6	33	32
7	62	65
8	80	72
9	80	80

Pruebe si existe diferencia con un nivel de significación de 1 por ciento.

Prueba de los signos para 2 muestras pareadas grandes

17. Una cadena de tiendas de autoservicio analiza la reestructuración de su área de cajas en una de sus sucursales para evaluar si aumenta el flujo de clientes. Los datos de 20 horas antes y después de la reestructuración arrojan 12 datos con signo de menos, 6 con signo más y 2 valores de 0. ¿La reestructuración de las cajas mejoró la afluencia de clientes? Realice la prueba con un nivel de significación de 0.05.

18. Para evaluar los resultados de un partido político en 2 elecciones sucesivas, se eligieron al azar 13 casillas de votación y se obtuvieron los siguientes resultados (en porcentaje de votos a favor):

Casilla	Primera elección	Segunda elección
1	65	61
2	73	70
3	39	40
4	28	27
5	50	54
6	33	29
7	96	76
8	62	58
9	14	23
10	45	43
11	90	87
12	82	80
13	46	49

Con un nivel de significación de 0.01, ¿se puede afirmar que existen evidencias de que el porcentaje de votos obtenidos por ese partido disminuyó de una elección a la otra?

19. En un estudio médico para evaluar si la participación de psiquiatras en el proceso de recuperación de alcohólicos mejoraba su patrón de conducta, se evaluó una muestra de 12 pacientes, al principio del tratamiento y al cabo de 2 meses, y se obtuvieron las siguientes calificaciones:

Paciente	Calificación inicial	Calificación final
1	31	32
2	40	39
3	41	42
4	88	99
5	57	73
6	56	57
7	65	60
8	35	45
9	68	71
10	71	70
11	80	80
12	84	85

Con un nivel de significación de 0.05, ¿se puede afirmar que existen evidencias de que la participación de los psiquiatras en el tratamiento de alcohólicos mejoró su patrón de conducta?

20. En una prueba de desgaste de llantas se evaluaron 11 tipos de neumáticos con 2 diseños de superficie de rodamiento diferentes y se evaluó el desgaste en una escala de 0 a 50, con los siguientes resultados:

Llanta	Diseño 1	Diseño 2
1	42	47
2	37	35
3	31	31
4	23	28
5	35	39
6	48	49
7	27	33
8	39	43
9	42	44
10	44	47
11	35	38

Evalúe si existe una diferencia entre los desgastes con los 2 diseños, con un nivel de significación de 5 por ciento.

21. Los datos siguientes corresponden a la presión sanguínea sistólica (el valor máximo de la tensión arterial cuando el corazón se contrae) de 13 personas, antes y después de un programa de ejercicio de una hora. Con un nivel de significación de 5%, ¿se puede afirmar que el programa de ejercicio ayuda a disminuir la presión sistólica?

Persona	Presión antes	Presión después
1	124	122
2	106	104
3	108	106
4	114	116
5	103	105
6	120	119
7	110	105
8	108	110
9	120	118
10	110	115
11	135	130
12	106	107
13	120	117

22. Los números de pasajeros que volaron entre las ciudades de México y Guadalajara en determinado vuelo de una aerolínea fueron (ida y vuelta, respectivamente): 222 y 179, 255 y 220, 239 y 226, 260 y 271, 245 y 239, 246 y 228, 260 y 248, 256 y 243, 239 y 241, 250 y 243, 247 y 244, y 249 y 259. Pruebe la hipótesis de que el mismo número de pasajeros viaja en ambas direcciones, con un nivel de significación de 1 por ciento.

17.4 Prueba de rangos con signo de Wilcoxon

Se utiliza para probar hipótesis nulas sobre la mediana, al igual que la prueba de los signos. También al igual que en dicha prueba, como se puede aplicar a una sola muestra o a 2 muestras apareadas (relacionadas), la hipótesis nula puede entonces referirse a 1 o a 2 medianas poblacionales. Esta prueba equivale a una prueba paramétrica sobre la media de una población, igual también que en la prueba de los signos, y se pueden realizar pruebas de 1 o de 2 extremos, como en la de los signos.

Sin embargo, y ahora sí a diferencia de la prueba de los signos, como la prueba de Wilcoxon considera la magnitud de la diferencia entre las mediana muestral y la media hipotética (1 muestra) o entre las 2 medianas muestrales (2 muestras apareadas), se trata de una prueba más sensible que la de los signos, pero, por esta misma razón, la escala de medición debe ser cuando menos de intervalo.

17.4.1 Características

Las principales características de esta prueba son:

1. Hipótesis nula: la prueba se hace sobre la o las medianas poblacionales.
2. Nivel de medición: la variable debe estar cuando menos en escala de intervalo.
3. Suposiciones: no se hacen suposiciones sobre la forma de la distribución de la variable en la población.
4. Extremos: se pueden hacer pruebas de 1 o de 2 extremos.
5. Número de muestras: se puede aplicar a una sola muestra o a 2 muestras relacionadas (apareadas).
6. Estadístico de prueba:

   a) Para muestras pequeñas ($n \leq 15$) se utiliza el estadístico T de Wilcoxon y la tabla 6 del apéndice.
   b) Para muestras grandes, cuando la muestra tiene cuando menos 16 elementos, se utiliza la aproximación normal, con z:

$$z = \frac{T - \mu_T}{\sigma_T} = \frac{p - \pi}{\sqrt{\frac{\pi Q}{n}}}$$

o con valores muestrales:

$$z = \frac{T - \mu_T}{s_T} \tag{17.7}$$

En donde

$$_T = \frac{n(n+1)}{4} \tag{17.8}$$

y

$$\sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{24}} \tag{17.9}$$

Como puede observarse, en ambos casos muestras grandes y muestras pequeñas, se calcula un estadístico muestral T, mediante el siguiente procedimiento:

I. Se calcula $d_i = (x_i - Med)$, la diferencia entre cada uno de los valores observados y la mediana que plantea la hipótesis nula (una muestra) o la diferencia entre cada par de valores apareados (dos muestras).
II. Si alguna de estas diferencias es 0 se le elimina, con la consiguiente reducción del tamaño de la muestra.
III. Se clasifican los valores absolutos de las diferencias restantes, de menor a mayor y se les asigna un rango, comenzando por 1 para la diferencia menor y continuando con las demás. Si hay empates en las diferencias absolutas se les asigna el promedio de los rangos correspondientes.
IV. Se suman, por separado, los rangos de las diferencias que hayan resultado positivas y de las negativas.

La menor de estas 2 sumas es el estadístico T de Wilcoxon para una prueba de 2 extremos. Para pruebas de un extremo, el estadístico T así obtenido se asocia con el sentido de la hipótesis nula. Si el valor calculado de la T es menor que el valor crítico se rechaza la hipótesis nula. En los ejemplos que siguen se revisan los detalles.

17.4.2 Prueba de rangos con signo de Wilcoxon para una muestra pequeña

Se ilustra este caso con el ejemplo siguiente.

Ejemplo 17.7

Una aseguradora estudia los registros de reclamaciones de seguros de daños automotrices y determinó que, en el pasado, la mediana de las reclamaciones fue de $8 500. Toma una muestra de 20 reclamaciones con los siguientes resultados:

14 000	3 500	9 000	12 000	2 500
5 000	15 000	8 000	8 000	16 000
9 200	31 000	7 500	10 000	6 800
6 500	20 200	8 700	4 700	6 000

Con un nivel de significación de 5%, ¿se afirmaría que hubo un cambio en la mediana de las reclamaciones de seguro de daños de automóviles?

Solución: Las hipótesis:

H_0: Med = \$8 500
H_1: $Med \neq$ \$8 500

En la tabla 17.8 se resumen los cálculos.

Tabla 17.8 Datos y cálculos para el ejemplo 17.7

	$d_i = (x_i - Med)$ $d_i = (x_i - 8\,500)$	Rango
14 000	5 500	15
5 000	–3 500	11.5
9 200	700	5
6 500	–2 000	9
3 500	–5 000	14
15 000	6 500	17
31 000	22 500	20
20 200	11 700	19
9 000	500	3
8 000	–500	3
7 500	–1 000	6
8 700	200	1
12 000	3 500	11.5
8 000	–500	3
10 000	1 500	7
4 700	–3 800	13
2 500	–6 000	16
16 000	7 500	18
6 800	–1 700	8
6 000	–2 500	10
Suma de rangos +		116.5
Suma de rangos –		93.5

Nótese que, como hay 3 diferencias iguales a 500 y los rangos que les corresponden son 2, 3 y 4, a esas 3 observaciones se les asigna 3, el promedio, y se continúa con el rango 5, que le corresponde a la diferencia absoluta de 700. También, como hay 2 diferencias absolutas de 3 500 y les corresponden los rangos 11 y 12, a ambas se les asigna el promedio, 11.5.

En la parte inferior de la tabla se anotan las sumas de los rangos positivos y negativos y, como la menor de estas sumas es la de los rangos negativos, el valor observado del estadístico T de Wilcoxon es 93.5.

Con referencia ahora a la tabla 6 del apéndice se puede ver que, para n = 20, una prueba de 2 extremos y α = 0.05, el valor crítico de T es 52 y como el valor observado de T fue de 93.5, no se rechaza la hipótesis nula y se concluye que no hubo cambio en la mediana de las reclamaciones de seguro de daños de automóviles.

17.4.3 Excel y la prueba de rangos con signo de Wilcoxon

Al llegar aquí, es posible que el lector ya hiciera uso de las capacidades de Excel para ayudarse con las operaciones necesarias para realizar este tipo de pruebas. Sin embargo, para no dejarlo pasar, se expone a continuación el procedimiento que se sugiere para hacerlo:

1. El primer paso consiste, por supuesto, en colocar los datos en una hoja de Excel. Utilizando los datos del ejemplo 17.7, este primer paso consistiría en colocar todos los datos en una sola columna, por ejemplo la columna A, empezando en la celda A1.
2. En segundo lugar, y para poder regresar al final del procedimiento a todos los datos a su orden original, se sugiere insertar una columna a la izquierda de los datos y llenarla con números consecutivos, con lo que los datos quedan en la columna B.
3. Para calcular las diferencias entre cada dato y la mediana supuesta en la hipótesis nula, se introduce la fórmula "=B1-8500", en la celda C1 y, con la cruz que aparece cuando se coloca el marcador del ratón en la esquina inferior derecha de esa celda B1, se "jala" esa fórmula hacia abajo hasta cubrir todos los renglones. Con esta operación ya se tienen en la columna C todas las diferencias.
4. Ahora se anota en la celda D1 la fórmula "=Abs(C1)" con lo que se tiene en esta celda el valor absoluto de la diferencia calculada en C1. También se "jala" esta fórmula hacia abajo hasta abarcar todos los renglones, tal como se acaba de describir en el paso 3.
5. Ahora es necesario asignar los rangos en orden de magnitud de las diferencias. Para hacer esto, se seleccionan todos los renglones de las columnas A, B, C y D y se abre la pestaña de "Datos" de la cinta de opciones de Excel y, aquí, se da clic en el cuadro de "Ordenar", con lo que aparece el siguiente cuadro de diálogo:

Si en este cuadro se da clic en la flecha que está a la derecha del espacio titulado "Ordenar por", aparece un listado con las 4 columnas: "Columna A", "Columna B", "Columna C" y "Columna D". Al elegir la columna D, que contiene los valores absolutos de los rangos, Excel automáticamente propone en el espacio de "Criterio de ordenación" que está del lado derecho, hacerlo "De menor a mayor". Al hacer clic en "Aceptar", Excel ordena todos los renglones de las 4 columnas desde el menor valor absoluto al mayor.

Ahora se anotan los rangos en la columna E. Para esto, es suficiente con anotar 1 y 2 en los renglones D1 y D2, marcar ambas celdas y jalar la cruz de la esquina inferior izquierda de esta selección con lo que se llenan de enteros consecutivos todos los renglones, es decir, los rangos.

Una vez que se tienen los rangos se deben revisar para reemplazar los rangos de las diferencias absolutas empatadas por el promedio de los rangos correspondientes.

Finalmente, para calcular la suma de estos rangos, se puede introducir en el renglón que sigue del último dato en la columna D, la fórmula "SUMA(…" e ir marcando con el ratón las celdas en donde están los rangos de las diferencias positivas, anotando una coma después de cada celda marcada.

Para los rangos positivos se termina con la fórmula de Excel "=SUMA(C1,C2,C5, C7,C12,C15,C17,C18,C19,C20)". Haciendo lo mismo con los rangos correspondientes a las diferencias negativas se obtienen las sumas de rangos que se muestran en la tabla 17.8.

En esta tabla se eliminaron todas las columnas intermedias de cálculo para dejar solamente las de los datos, las diferencias y los rangos. Se regresaron los datos al orden original.

NOTA

Es importante asegurarse que no esté palomeada la opción de "Mis datos tienen encabezados" que aparece en la esquina superior derecha de este cuadro de diálogo de Excel.

NOTA

La operación de jalar la cruz hacia abajo se puede simplificar si, una vez que aparece, simplemente se da doble clic porque, con esto, Excel simplemente llena todos los renglones.

17.4.4 Prueba de rangos con signo de Wilcoxon para una muestra grande (aproximación normal)

En el ejemplo siguiente se ilustra el procedimiento para llevar a cabo esta prueba de hipótesis de rangos con signo de Wilcoxon para una muestra grande, mediante una aproximación con la distribución normal.

Ejemplo 17.8

Las siguientes son las calificaciones obtenidas en una muestra aleatoria de 35 estudiantes de licenciatura en una encuesta desarrollada por el departamento de administración escolar:

7.50	6.48	7.41	7.33	9.22
8.42	7.82	7.71	8.32	8.65
7.25	8.00	8.60	6.73	8.05
6.00	9.56	7.76	8.24	6.25
7.65	6.05	9.15	9.04	8.10
8.55	6.81	9.40	8.48	6.10
6.96	7.07	6.90	7.14	7.88

La calificación mediana histórica es de 7.4 y se desea probar, con un nivel de significación si la calificación mediana actual subió o no, con un nivel de significación $\alpha = 0.01$.

Solución: Las hipótesis:

$$H_0: Med_1 \leq Med_2$$
$$H_1: Med_1 > Med_2$$

Con un nivel de significación de $\alpha = 0.01$ y una prueba de un extremo, el valor crítico de z es 2.33, ya que $P(z \geq 2.33) = 0.01$.

En la tabla 17.9 se muestran los datos y las operaciones realizadas para encontrar las sumas de rangos y el estadístico T de Wilcoxon.

Tabla 17.9 Datos y operaciones para el ejemplo 17.8

7.50	0.10	3
8.42	1.02	22
7.25	–0.15	4
6.00	–1.40	30
7.65	0.25	5
8.55	1.15	24.5
6.96	–0.44	11
6.48	–0.92	20.5
7.82	0.42	10
8.00	0.60	15
9.56	2.16	35
6.05	–1.35	29
6.81	–0.59	14
7.07	–0.33	8
7.41	0.01	1

(*continúa*)

Tabla 17.9 (*continuación*)

7.71	0.31	7
8.60	1.20	26
7.76	0.36	9
9.15	1.75	32
9.40	2.00	34
6.90	–0.50	13
7.33	–0.07	2
8.32	0.92	21.5
6.73	–0.67	17
8.24	0.84	19
9.04	1.64	31
8.48	1.08	23
7.14	–0.26	6
9.22	1.82	33
8.65	1.25	27
8.05	0.65	16
6.25	–1.15	24.5
8.10	0.70	18
6.10	–1.30	28
7.88	0.48	12
Suma de rangos +		424
Suma de rangos –		207

Ahora, como la muestra tiene 35 elementos, se puede utilizar la aproximación normal con:

$$_T = \frac{n(n+1)}{4} = \frac{35(36)}{4} = 315$$

y

$$\sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{24}} = \sqrt{\frac{35(36)(71)}{24}} = \sqrt{3\,727.5} = 61.05$$

Y, el estadístico de prueba calculado:

$$z = \frac{T - \mu_T}{S_T} = \frac{424 - 315}{61.05} = 1.785$$

Como el estadístico de prueba crítico es $z = 2.33$, y ese valor calculado es menor, no se rechaza la hipótesis nula y se concluye que la mediana de las calificaciones de esa encuesta no ha cambiado.

17.4.5 Prueba de rangos con signo de Wilcoxon para 2 muestras apareadas pequeñas

En el ejemplo siguiente se ilustra el procedimiento para llevar a cabo esta prueba de hipótesis de rangos con signo de Wilcoxon para dos muestras apareadas pequeñas.

Ejemplo 17.9

En una investigación de mercado se pidió a una muestra de 15 consumidores que evaluaran 2 marcas de café instantáneo mediante un sistema de puntaje que combina diversos criterios. El investigador desea probar si existe una preferencia por la marca 1, con un nivel de significación de 1%. En la tabla 17.10 se muestran los datos, junto con los cálculos necesarios.

Tabla 17.10 Datos y operaciones para el ejemplo 17.9

Consumidor	Marca 1	Marca 2	Diferencia	Rango
1	25	20	5	8
2	29	31	–2	3
3	33	23	10	13.5
4	29	23	6	9
5	25	24	1	1
6	34	26	8	11
7	24	28	–4	6.5
8	32	28	4	6.5
9	25	27	–2	3
10	35	25	10	13.5
11	23	23	0	
12	33	26	7	10
13	31	22	9	12
14	29	31	–2	3
15	25	28	–3	5
		Suma de rangos +		84.5
		Suma de rangos –		20.5

Solución: Las hipótesis:

$$H_0: Med_1 \leq Med_2$$
$$H_1: Med_1 > Med_2$$

Revisando los cálculos de la tabla 17.10 se puede ver que el valor empírico del estadístico T es igual a 20.5, que es la menor de las 2 sumas de rangos con signo; revisando ahora la tabla 6 del apéndice se puede ver que, para $n = 14$ (se eliminó una observación cuya diferencia fue 0), una prueba de un extremo y $\alpha = 0.01$, el valor crítico de T es 16 y como el valor observado de T fue de 20.5, no se rechaza la hipótesis nula y se concluye que no hay diferencia entre los consumidores respecto a las 2 marcas de café.

17.4.6 Prueba de rangos con signo de Wilcoxon para 2 muestras apareadas grandes (aproximación normal)

En el ejemplo siguiente se ilustra el procedimiento para llevar a cabo esta prueba de hipótesis de rangos con signo de Wilcoxon para dos muestras apareadas grandes, mediante una aproximación con la distribución normal.

Ejemplo 17.10

Un inventor afirma que ideó un aditivo para gasolina que aumenta el rendimiento del combustible. Un posible financiador del producto hace una prueba con 18 automóviles diferentes que se conducen en las mismas condiciones, con y sin el aditivo, y obtiene los resultados que se muestran en la tabla 17.11. Probar, con un nivel de significación de 0.05, si se puede aceptar la afirmación del inventor como verdadera.

Tabla 17.11 Datos para el ejemplo 17.10

Rendimiento de la gasolina (km × L)	
Con aditivo	**Sin aditivo**
24.46	25.64
38.34	38.10
35.52	37.16
21.64	23.52
24.23	23.99
19.76	18.58
22.82	22.58
20.23	23.29
25.87	27.99
31.05	30.58
42.57	42.57
17.64	19.05
22.35	23.05
25.64	25.64
20.46	24.23
35.52	38.10
31.52	30.58
28.93	32.46

Solución: Las hipótesis:

H_0: $Med_1 \leq Med_2$
H_1: $Med_1 > Med_2$

Con un nivel de significación de $\alpha = 0.05$ y una prueba de un extremo, el valor crítico de z es 1.645, ya que $P(z \geq 1.645) = 0.05$. En la tabla 17.12 se muestran los datos y las operaciones realizadas para encontrar las sumas de rangos y el estadístico T de Wilcoxon.

Tabla 17.12 Datos y operaciones para el ejemplo 17.10

	Rendimiento de la gasolina (km × L)			
	Con aditivo	**Sin aditivo**	**Diferencias**	**Rangos**
1	25.64	24.46	1.18	7.5
2	38.1	38.34	–0.24	2
3	37.16	35.52	1.64	10
4	23.52	21.64	1.88	11
5	23.99	24.23	–0.24	2
6	18.58	19.76	–1.18	7.5
7	22.58	22.82	–0.24	2
8	23.29	20.23	3.06	14
9	27.99	25.87	2.12	12
10	30.58	31.05	–0.47	4
11	42.57	42.57	0	
12	19.05	17.64	1.41	9
13	23.05	22.35	0.7	5
14	25.64	25.64	0	
15	24.23	20.46	3.77	16
16	38.1	35.52	2.58	13
17	30.58	31.52	–0.94	6
18	32.46	28.93	3.53	15
		Suma de rangos +		112.5
		Suma de rangos –		23.5

Con una muestra de 16 elementos (no se toman en cuenta las 2 observaciones cuya diferencia fue 0) se utiliza la aproximación normal:

$$_T = \frac{n(n+1)}{4} = \frac{16(17)}{4} = 68$$

y

$$\sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{24}} = \sqrt{\frac{16(17)(33)}{24}} = \sqrt{874} = 19.34$$

El estadístico de prueba calculado:

$$z = \frac{T - \mu_T}{S_T} = \frac{112.5 - 68}{19.34} = 2.30$$

Como el estadístico de prueba crítico es $z = 1.645$, y ese valor calculado es mayor, se rechaza la hipótesis nula y se concluye que la mediana del rendimiento de los autos que utilizaron gasolina con aditivo es mayor que en los que no lo utilizaron.

ejercicios 17.4 Prueba de rangos con signo de Wilcoxon

1. El ejemplo 17.7 del texto se resolvió asumiendo que se trataba de una muestra pequeña, aun cuando el tamaño de ésta era $n = 20$. Resuélvalo ahora utilizando la aproximación normal y verifique si se llega a la misma conclusión.
2. El ejemplo 17.8 se resolvió con el método de la aproximación normal, puesto que el tamaño de la muestra era $n = 35$. Resuélvalo ahora utilizando el método exacto (que se utiliza para muestras pequeñas) y verifique si se llega a la misma conclusión.
3. El ejemplo 17.9 se resolvió con el método aplicable a muestras pequeñas. Resuélvalo ahora con la aproximación normal y verifique si se llega a la misma conclusión. Comente los resultados.
4. El ejemplo 17.10 se resolvió con el método de la aproximación normal, con un tamaño de muestra que era $n = 16$. Resuélvalo ahora utilizando el método exacto (que se utiliza para muestras pequeñas) y verifique si se llega a la misma conclusión.

Prueba de rangos con signo de Wilcoxon para una muestra pequeña

5. De los ejercicios de la sección 17.3 "La prueba de los signos", resuelva los ejercicios 1 a 5 utilizando esta prueba de Wilcoxon.

Prueba de rangos con signo de Wilcoxon para una muestra grande (aproximación normal)

6. De los ejercicios de la sección 17.3 "La prueba de los signos", resuelva los puntos 6 a 11 utilizando esta prueba de Wilcoxon.

Prueba de rangos con signo de Wilcoxon para 2 muestras apareadas pequeñas

7. De los ejercicios de la sección 17.3 "La prueba de los signos", resuelva los ejercicios 12 a 16 utilizando esta prueba de Wilcoxon.

Prueba de rangos con signo de Wilcoxon para 2 muestras apareadas grandes (aproximación normal)

8. De los ejercicios de la sección 17.3 "La prueba de los signos", resuelva los puntos 17 a 22 utilizando esta prueba de Wilcoxon.

17.5 Prueba *U* de Mann-Whitney para 2 muestras independientes

Esta prueba se utiliza para determinar si 2 muestras se obtuvieron de la misma población y la hipótesis nula es de igualdad de medias con los resultados obtenidos de 2 muestras aleatorias independientes.

Es equivalente a la prueba sobre 2 medias que se estudió en el capítulo 11 pero, en este caso de Mann-Whitney, y a diferencia de la primera, no es necesario suponer que las 2 poblaciones tienen la misma varianza ni la misma distribución.

Esta prueba de Mann-Whitney es una buena opción ante la prueba paramétrica sobre la diferencia entre 2 medias; se hace utilizando la t de Student como estadístico de prueba, si no se desea o no se pueden cumplir los supuestos en los que se basa o si la escala de medición de la variable es menor a intervalo.

17.5.1 Características

Las principales características de esta prueba de hipótesis son:

1. Hipótesis nula: la prueba se hace sobre las medias poblacionales.
2. Nivel de medición: la variable debe estar cuando menos en escala ordinal.
3. Suposiciones: no se hacen suposiciones
4. Extremos: se pueden hacer pruebas de 1 o de 2 extremos.
5. Número de muestras: dos, independientes.
6. Estadístico de prueba:

   a) Para muestras pequeñas ($n \leq 20$) se utiliza el estadístico U de Mann-Whitney y la tabla 7 del apéndice.

El estadístico U de Mann-Whitney tiene 2 formas, y se utiliza una de ellas, según se trate de una prueba de 1 o de 2 extremos:

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 \tag{17.10}$$

y

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2 \tag{17.11}$$

en donde R_1 y R_2 son las sumas de rangos para las muestras 1 y 2.

Si la hipótesis nula es cierta, la suma de los rangos de las 2 muestras debe ser aproximadamente igual. Por ello, en una prueba de 2 colas se utiliza la U que sea menor y se le compara con los valores críticos de la tabla 7 del apéndice. Si la U observada que es más pequeña es menor o igual que la U crítica, la de las tablas, entonces se rechaza la hipótesis nula. En una prueba de un extremo se rechaza la hipótesis nula si U_1 es menor o igual que la U crítica.

b) Para muestras grandes, cuando la muestra tiene cuando menos 16 elementos, se utiliza la aproximación normal, con z:

$$z = \frac{U_1 - \mu_{U_1}}{\sigma_{U_1}} \tag{17.12}$$

en donde,

$$_{U_1} = \frac{n_1 n_2}{2} \tag{17.13}$$

y

$$\sigma_{U_1} = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} \tag{17.14}$$

El procedimiento para calcular las 2 U es:

1. Combinar los valores de las 2 muestras en un arreglo ordenado, de menor a mayor, identificando a cuál de las 2 muestras pertenece cada valor.
2. Asignar rangos, de 1 en adelante, a partir del menor valor. Cuando se dan empates en los rangos se asigna el promedio.
3. Se suman los rangos de las 2 muestras.

Ejemplo 17.11

Se desea probar si la resistencia de cierto tipo de cable de cobre es la misma para 2 tratamientos de acabado. Se seleccionan al azar 2 muestras de 10 tramos de cable de cada tipo y se obtienen las resistencias que se muestran en la tabla 17.13. Probar la hipótesis de que no existe diferencia entre las distribuciones de las resistencias en los 2 tipos de cable, con un nivel de significación de 0.01.

Tabla 17.13 Datos del ejemplo 17.11

Acabado	
1	**2**
3.21	3.49
3.43	3.37
3.35	3.67
3.51	3.5
3.39	3.31
3.17	3.29
3.48	3.52
3.42	3.37
3.29	3.44
3.4	3.53

Solución: Las hipótesis:

H_0: $\mu_1 = \mu_2$
H_1: $\mu_1 \neq \mu_2$

En la tabla 7 del apéndice se puede ver que el valor crítico del estadístico de prueba U, para un nivel de significación de 0.01 y prueba de 2 extremos es 19 $\left(\alpha = \frac{0.01}{2} = 0.005\right)$.

Ahora, en la tabla 17.74 se muestran los datos ordenados para las 2 muestras, las asignaciones de rangos y las sumas de los rangos de cada muestra.

Tabla 17.14 Datos y operaciones para el ejemplo 17.11

a	3.17	1
a	3.21	2
a	3.29	3.5
b	3.29	3.5
b	3.31	5
a	3.35	6
b	3.37	7.5
b	3.37	7.5
a	3.39	9
a	3.4	10
a	3.42	11
a	3.43	12
b	3.44	13
a	3.48	14
b	3.49	15
b	3.5	16
a	3.51	17
b	3.52	18
b	3.53	19
b	3.67	20
	Suma 1	85.5
	Suma 2	124.5

Para llegar a esta tabla, primero se identificaron los datos de la muestra con "a" y los de la muestra 2 con "b", para después reunirlos todos en una sola columna. Una vez teniendo una sola columna se les ordena de menor a mayor y se asignan rangos, comenzando con 1 y asignando el rango promedio en caso de empates. Se tiene, entonces, que la $R_1 = 85.5$ y $R_2 = 124.5$.

Los estadísticos de prueba calculados:

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 = 10(10) + \frac{100(11)}{2} - 85.5 = 69.5$$

y

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2 = 10(10) + \frac{10(11)}{2} - 124.5 = 30.5$$

Por lo que el estadístico de prueba calculado o empírico es $U = 30.5$. Como el valor crítico que se determinó antes es 19 y la U calculada es mayor que 19, no es posible rechazar la hipótesis nula y se concluye que no hay elementos para pensar que hay diferencia entre las resistencias de los cables con los 2 distintos acabados.

EJEMPLO 17.12

Resolviendo de nuevo el ejemplo 17.11 con la aproximación normal:

$$_{U_1} = \frac{n_1\ n_2}{2} = \frac{10(10)}{2} = 50$$

y

$$\sigma_{U_1} = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} = \sqrt{\frac{10(10)(21)}{12}} = \sqrt{175} = 13.23$$

$$z = \frac{U_1 - \mu_{U_1}}{\sigma_{U_1}} = \frac{69.5 - 50}{13.23} = 1.47$$

Para un nivel de significación de 0.01 y una prueba de 2 colas, el valor crítico de z es 2.57, ya que $P(-2.57 \le z \le 2.57)$. Como el valor calculado de la z, 1.47, está entre –2.57 y 2.57, no se rechaza la hipótesis nula y se concluye, igual que antes, que no hay elementos para pensar que hay diferencia entre las resistencias de los cables con los 2 distintos acabados.

EJEMPLO 17.13

Se prueba una dieta que se afirma aumenta el peso de las aves en mayor medida que la dieta estándar. Se compararon los resultados con otra muestra aleatoria de 15 guajolotes bajo la dieta estándar y se obtuvieron los siguientes pesos, en kilogramos:

Dieta estándar 1	Dieta de prueba 2
7.39	9.66
4.58	10.80
4.85	6.99
6.12	8.89
6.76	5.44
5.35	6.30
6.49	8.53
4.63	8.71
5.44	6.94
6.67	9.12
10.70	6.71
6.85	8.57
6.58	9.39
8.35	9.57
5.99	7.17

Probar, con un nivel de significación de 0.05 si la dieta de prueba efectivamente hace que los guajolotes suban más de peso que con la dieta estándar.

Solución: Las hipótesis:

$$H_0 : \mu_1 \geq \mu_2$$
$$H_1 : \mu_1 < \mu_2$$

En la tabla 17.15 se muestran los datos y las operaciones para calcular las sumas de rangos.

Tabla 17.15 Datos y operaciones para el ejemplo 17.13

a	4.58	1
a	4.63	2
a	4.85	3
a	5.35	4
a	5.44	5.5
b	5.44	5.5
a	5.99	7
a	6.12	8
b	6.3	9
a	6.49	10
a	6.58	11
a	6.67	12
b	6.71	13
a	6.76	14
a	6.85	15
b	6.94	16
b	6.99	17
b	7.17	18
a	7.39	19
a	8.35	20
b	8.53	21
b	8.57	22
b	8.71	23
b	8.89	24
b	9.12	25
b	9.39	26
b	9.57	27
b	9.66	28
a	10.7	29
b	10.8	30
Suma de rangos 1		160.5
Suma de rangos 2		304.5

Como se trata de muestras de tamaño 15 se puede utilizar la aproximación normal y, entonces, el valor crítico del estadístico de prueba, z, para una prueba de un extremo, es 1.645, ya que $P(z \geq 1.645) = 0.05$.

$$_{U_1} = \frac{n_1 n_2}{2} = \frac{15(15)}{2} = 112.5$$

y

$$\sigma_{U_1} = \sqrt{\frac{n_1\, n_2 (n_1 + n_2 + 1)}{12}} = \sqrt{\frac{15(15)(31)}{12}} = \sqrt{581.25} = 24.11$$

$$z = \frac{U_1 - \mu_{U_1}}{\sigma_{U_1}} = \frac{160.5 - 112.5}{24.11} = 1.99$$

Como el valor crítico de z es 1.645 y el valor calculado de la z, 1.99 es mayor, se rechaza la hipótesis nula y se concluye que existe evidencia para afirmar que, efectivamente, la dieta que se propone aumenta más de peso a los guajolotes que la dieta estándar.

EJERCICIOS 17.5 Prueba *U* de Mann-Whitney para 2 muestras independientes

1. Se eligieron al azar 10 familias de 4 personas (2 adultos y 2 niños de menos de 15 años de edad) en 2 colonias distintas de la Ciudad de México. Se investigó cuánto gastaron en alimentación un día de la semana anterior y se obtuvieron los siguientes datos:

Colonia 1 X_1	Colonia 2 X_2
224.34	240.36
247.80	193.89
215.67	257.73
222.90	276.48
228.00	283.77
204.57	178.05
271.35	178.28
225.45	294.57
215.85	197.25
238.14	266.16

Suponiendo que la semana considerada es representativa de los gastos diarios en alimentación para estas familias, ¿existe una diferencia real entre los gastos semanales en alimentación para las familias de la colonia 1 y las familias de la colonia 2, a un nivel de significación de $\alpha = 0.05$?

2. Se estudiaron los precios al menudeo de cierto producto en las ciudades de Monterrey y Guadalajara. Se eligieron

al azar 10 establecimientos comerciales en Guadalajara y 12 en Monterrey y se obtuvieron los siguientes resultados:

Guadalajara	Monterrey
159	144
171	141
186	168
165	165
153	156
147	171
156	159
174	150
165	147
171	162
	156
	180

¿Se puede afirmar, a un nivel de significación del 0.05, que existe una diferencia en los precios de estos productos en Monterrey y Guadalajara?

3. Se pidió a 2 grupos de economistas, uno del sector público y otro del sector privado, que hicieran un pronóstico sobre el posible crecimiento de la economía, en términos del Producto Interno Bruto (PIB), para el año siguiente y se obtuvieron los pronósticos que se muestran a continuación:

Pronóstico del crecimiento del PIB al año siguiente (%)	
Economista del sector público	Economista del sector privado
4.5	3.2
5.9	4.9
4.0	2.4
6.7	5.7
6.4	0.1
8.5	2.9
8.8	3.0
7.6	

Con un nivel de significación de 0.05, pruebe la hipótesis de que los pronósticos de los economistas del sector público tienden a ser mayores que los de los economistas del sector privado.

4. Se realizó una prueba de mercado en grupos de hombres y mujeres acerca de su preferencia sobre una bebida energética que está en proceso de desarrollo para su lanzamiento al mercado. Se les pidió a los miembros de ambos grupos que calificaran la bebida en una escala de 0 a 10 y se obtuvieron los resultados siguientes:

Calificación de la bebida	
Mujeres	Hombres
8	2
6	10
7	7
4	6
6	6
10	8
8	6
6	5

Pruebe la hipótesis de que no existe diferencia entre las preferencias por la bebida entre los 2 sexos, con un nivel de significación de $\alpha = 0.01$.

5. Se tomaron muestras de los tipos de cambio de venta de dólares en diversas casas de cambio de los aeropuertos de la Ciudad de México y de diversas ciudades fronterizas mexicanas, y se obtuvieron los siguientes resultados:

Tipos de cambio peso-dólar en los aeropuertos de	
Ciudad de México	Ciudades fronterizas mexicanas
12.40	12.60
12.45	12.45
12.30	12.50
12.25	12.75
12.80	12.30
12.55	12.25
12.43	12.35
12.67	12.40

Pruebe si los precios de venta de dólares son diferentes en las casas de cambio del aeropuerto de la Ciudad de México, en comparación con las casas de cambio en aeropuertos de ciudades fronterizas mexicanas, con un nivel de significación de $\alpha = 0.05$.

17.6 Prueba de suma de rangos de Kruskal-Wallis para más de 2 medias

La prueba de la suma de rangos de Kruskal-Wallis se utiliza para probar la igualdad de más de 2 medias y es el equivalente no paramétrico de la prueba de análisis de varianza de un factor, pero en el caso de esta prueba de Kruskal-Wallis, no se requiere la suposición de que las muestras provienen de poblaciones en las que la variable se distribuye de forma aproximadamente normal.

El procedimiento para realizar la prueba es similar al que se vio en la sección anterior para la prueba U de Mann-Whitney para 2 muestras: considerando a todas las observaciones de todas las muestras como si fueran una sola, ordenarlas de menor a mayor y asignarles rango, de menor a mayor, comenzando con 1 y promediando los rangos de las observaciones repetidas. Finalmente, se calculan las sumas de los rangos de cada muestra para calcular el estadístico de prueba H:

$$H = \left(\frac{12}{n(n+1)}\right)\left(\sum_{i=1}^{k} \frac{R_1^2}{n_i}\right)\left(-3(n+1)\right) \qquad \textbf{(17.15)}$$

Entre más grandes sean las diferencias entre las medias de las k muestras, mayor será el valor de H. Por ello, se rechaza la hipótesis nula de igualdad de medias para valores grandes de H.

Cuando el número de muestras es de cuando menos 3 y también se verifica que al menos una de las muestras tiene por lo menos 5 elementos, el estadístico teórico de prueba es la χ^2 (ji cuadrada) que se estudió detalladamente en el capítulo 11. Para esta prueba, los grados de libertad son $k - 1$, el número de muestras menos 1.

Ejemplo 17.14

A un grupo de trabajadores de nuevo ingreso a una planta fabril se les asigna al azar a uno de 3 grupos de capacitación que utilizan otros tantos métodos distintos. Al final del proceso de capacitación se les evalúa y se obtienen los resultados que se muestran a continuación:

Método de capacitación		
A	**B**	**C**
95	86	90
88	83	68
93	80	73
75	85	77
87	62	70
98	73	80
85	81	
	84	

Utilizar la prueba de Kruskal-Wallis para probar si los 3 métodos de capacitación son igualmente efectivos. Utilizar en $\alpha = 0.01$.

Solución: Las hipótesis:

H_0: $\mu_1 = \mu_2 = \mu_3$

H_1: cuando menos una de las medias no es igual a las otras.

En la tabla 17.16 se muestran las operaciones para el cálculo de los totales de rangos.

Tabla 17.16 Operaciones para el ejemplo 17.14

b	62	1
c	68	2
c	70	3
b	73	4.5
c	73	4.5
a	75	6
c	77	7
b	80	8.5
c	80	8.5
b	81	10
b	83	11
b	84	12
a	85	13.5
b	85	13.5
b	86	15
a	87	16
a	88	17
c	90	18
a	93	19
a	95	20
a	98	21
Suma de rangos A		112.5
Suma de rangos B		75.5
Suma de rangos C		43

El valor calculado del estadístico de prueba:

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1)$$

$$= \frac{12}{21(22)}\left(\frac{112.5^2}{7} + \frac{75.5^2}{8} + \frac{43^2}{6}\right) - \left(3(21+1)\right)$$

$$= \left(0.025974(1\,808.04 + 712.53 + 308.17)\right) - 66 = 7.474$$

El valor crítico del estadístico de prueba, con $k - 2$ grados de libertad y $\alpha = 0.01$ es 9.21, ya que $P(x^2 \geq 9.21 | gl = 2) = 0.01$.

Como el valor empírico (calculado) del estadístico de prueba es inferior al valor crítico, se acepta la hipótesis nula y se concluye que todas las medias son iguales, en términos del planteamiento del ejemplo, todos los métodos de capacitación producen un rendimiento igual o similar.

ejercicios 17.6 Prueba de suma de rangos de Kruskal-Wallis para más de 2 medias

1. Para evaluar los avances en un programa de uso obligatorio de transporte escolar colectivo, se dividió la ciudad en sectores con aproximadamente la misma cantidad de escuelas y se tomaron muestras aleatorias de escuelas en 4 sectores para investigar cuántas de ellas se incorporaron ya al programa, con los resultados que se muestran a continuación:

Sector 1	Sector 2	Sector 3	Sector 4
70	40	60	10
15	25	25	20
25	20	40	5
40	55	45	30
20	50	30	5
		65	

¿Se puede decir que el avance en la incorporación al programa de transporte escolar colectivo obligatorio es igual en todos los sectores? Realice la prueba con un nivel de significación de 6 por ciento.

2. Se desea evaluar si existen diferencias significativas en los gastos de transporte en los que incurren los habitantes de áreas urbanas, suburbanas y rurales. Se toman muestras de habitantes de estas 3 áreas en diferentes regiones del país y se pide información en porcentaje sobre los ingresos familiares totales. Con los datos que aparecen en la tabla siguiente, pruebe la hipótesis de que no existe diferencia entre los porcentajes del ingreso familiar que los habitantes de las 3 zonas dedican al transporte, con un nivel de significación de $\alpha = 0.01$.

Zona urbana	Zona suburbana	Zona rural
5.2	6.0	4.4
4.9	6.8	5.3
4.0	7.7	6.3
6.3	5.0	5.7
4.3	5.3	3.9
4.2	6.9	5.9
5.5		4.8

3. Para comparar los salarios de profesores universitarios en distintos estados de la República, se tomaron muestras de profesores con calificaciones y experiencia similares en universidades privadas y se obtuvieron los siguientes resultados:

Salarios en las universidades del estado 1	Salarios en las universidades del estado 2	Salarios en las universidades del estado 3
31 000	20 200	27 000
25 500	20 300	32 000
28 000	20 600	36 000
25 000	20 900	36 000
22 900	22 200	35 200
26 700	20 300	35 300
29 300	21 200	36 200

Pruebe la hipótesis de que no existe diferencia entre los salarios de los profesores en las universidades de los 3 estados, con un nivel de significación de $\alpha = 0.05$.

4. Una asociación de dietistas, de nombre IIMAS, desea probar si es necesario distinguir entre 4 variantes de su "dieta milagrosa", a partir de los resultados en reducción de peso de 25 mujeres con sobrepeso, a quienes se les administraron al azar las variantes, y con las siguientes reducciones de peso, después de 3 semanas de dieta:

Variante 1	Variante 2	Variante 3	Variante 4
2.57	2.01	2.48	2.02
2.05	2.31	2.65	2.76
2.06	1.97	1.87	2.07
1.64	1.85	2.20	2.04
1.90		2.12	2.3
2.14		2.15	2.27
2.28		2.81	
2.03			

Realice la prueba con nivel de significación de $\alpha = 0.05$.

5. Se obtuvieron datos de 21 estudiantes a quienes se les aplicaron al azar 3 técnicas diferentes para aprender a programar computadoras en lenguaje C++ y, después del entrenamiento, se les aplicaron pruebas para evaluar sus capacidades como programadores y se obtuvieron los siguientes resultados:

Técnica 1	Técnica 2	Técnica 3
76	77	75
79	81	74
73	80	77
74	78	72
78	79	74
76	82	70
	80	71
	83	

¿Es posible concluir, con un nivel de significación de 1% que no existen diferencias en los resultados de las 3 técnicas en términos de los estudiantes capacitados? Utilice $\alpha = 0.01$.

17.7 Prueba de Friedman para diseños en bloques aleatorizados

Esta prueba es el equivalente no paramétrico del análisis de varianza de 2 factores que se vio en la sección 12.7 del capítulo 12, siendo la principal diferencia que las muestras para la prueba paramétrica deben ser independientes en tanto que, en esta prueba de Friedman se trata de muestras dependientes, muestras a las que se les aplicaron los distintos tratamientos.

Otra diferencia consiste en que, en el análisis de varianza de 2 factores, se prueba si existen diferencias tanto entre tratamientos (columnas) como entre bloques (renglones), mientras que en esta prueba de Friedman sólo se prueba si existen diferencias entre los tratamientos. Por ello, el procedimiento implica el cálculo de la suma de rangos para las observaciones en los distintos tratamientos; es decir, los rangos se asignan a las observaciones de los diferentes tratamientos que, de acuerdo con la disposición acostumbrada de éstos, se utilizan en los rangos por renglón. Esta prueba fue propuesta por Milton Friedman, economista que ganó el premio Nobel de su especialidad en 1976. El estadístico de prueba es:

$$F_r = \frac{12}{nk(k+1)} \sum_{j=1}^{k} R_j^2 - 3n(k+1) \qquad \textbf{(17.16)}$$

en donde:

n es el número de bloques (renglones).
k es el número de tratamientos (columnas).
R_j es la suma de los rangos del j-ésimo tratamiento (columna).

Con muestras de cuando menos 5 elementos, este estadístico F_r sigue una distribución χ^2 con $k - 1$ grados de libertad.

Ejemplo 17.15

Se les aplica a 10 estudiantes de secundaria elegidos al azar una prueba psicológica para medir agresividad ante 3 programas de televisión medidos con niveles de agresividad baja, media y alta, con puntuaciones de 60 a 100 y se obtienen los resultados que se muestran en la tabla 17.17.

Tabla 17.17 Resultados de la prueba de agresividad

	Niveles de agresividad en los programas de TV		
Estudiante	**Baja**	**Media**	**Alta**
1	63 (1)	70 (2)	72 (3)
2	81 (1)	85 (3)	83 (2)
3	76 (2)	75 (1)	79 (3)
4	68 (1)	69 (2)	75 (3)
5	79 (1)	81 (2)	87 (3)
6	65 (1)	68 (3)	67 (2)
7	78 (1)	86 (2)	91 (3)
8	80 (1)	87 (2)	89 (3)
9	85 (2)	86 (3)	82 (1)
10	69 (1)	74 (2)	78 (3)
Suma de rangos	12	22	26
Cuadrado de la suma de rangos	144	484	676

Probar que la respuesta de agresividad es igual ante los 3 programas de televisión, con nivel de significación de 1 por ciento.

Solución: Las hipótesis:

H_0: $\mu_1 = \mu_2 = \mu_3$
H_1: Cuando menos una de las medias no es igual a las otras.

En la misma tabla 17.17 se anotan los rangos de cada renglón entre paréntesis y en los últimos renglones se anotan las sumas de rangos de tratamientos y sus correspondientes cuadrados.

Sustituyendo ahora en la fórmula:

$$F_r = \frac{12}{nk(k+1)} \sum_{j=1}^{k} R_j^2 - 3n(k+1)$$

$$= \frac{12}{10(3)(4)}(144 + 484 + 676) - 3(10)(4)$$

$$= 0.1(1\,304) - 120 = 130.4 - 120 = 10.4$$

El valor crítico del estadístico de prueba χ^2 es 9.21, ya que $(\chi^2 \geq 9.21 \mid gl = 2) = 0.01$.

Como el valor empírico (calculado) del estadístico de prueba es mayor de 9.21, se rechaza la hipótesis nula y se concluye que cuando menos una de las medias no es igual a las otras.

ejercicios 17.7 Prueba de Friedman para diseños en bloques aleatorizados

1. En el departamento de investigación y desarrollo de una empresa de computadoras se desea probar si existen diferencias entre los modelos de computadoras personales. Para tal efecto, se les pidió a 9 ingenieros del departamento de producción que clasificaran 3 modelos de computadoras en orden de preferencia y se les asignó al azar a cada una de las máquinas y se les pidió que calificaran con 1 a la de mayor preferencia, con 3 a la de menor y con 2 a la intermedia. Los datos que se obtuvieron son:

Ingeniero	Modelo I	Modelo II	Modelo III
1	2	3	1
2	2	3	1
3	2	3	1
4	1	3	2
5	3	2	1
6	1	2	3
7	2	3	1
8	1	3	2
9	1	3	2

Realice la prueba con un nivel de significación de $\alpha = 0.05$.

2. Se formaron 3 grupos de 8 personas con problemas de anorexia para evaluar 3 tipos de tratamiento para aumentar de peso que incluían seguimiento psicológico, régimen alimenticio y una serie de ejercicios. Los 3 grupos se conformaban de acuerdo con su sexo, tamaño y estructura ósea y a una persona de cada grupo se le asignó uno de los regímenes. En la tabla que sigue se muestran los resultados de aumento de peso, después de 3 meses de tratamiento:

Bloque	Régimen A	Régimen B	Régimen C
1	7.5	6.9	8.8
2	9.2	10.2	12.0
3	5.1	4.4	5.2
4	7.1	6.1	7.2
5	6.9	8.3	8.0
6	8.5	7.3	7.6
7	8.1	7.3	7.5
8	6.3	6.0	6.1

Utilice la prueba de Friedman para determinar si existe diferencia entre los aumentos de peso debidos a los distintos tratamientos. Utilice un nivel de significación de 0.01.

3. Se trataba de evaluar si existen diferencias entre los avalúos realizados por 3 expertos valuadores de bienes raíces. Para ello, se les pidió evaluar 7 bienes raíces elegidos al azar y sus evaluaciones (en millones de pesos) fueron las siguientes:

Bien raíz	Valuador 1	Valuador 2	Valuador 3
1	5.7	6.2	6.4
2	6.4	6.1	6.2
3	7.5	8.0	7.6
4	8.0	8.2	8.1
5	8.9	9.1	8.5
6	22.4	27.5	2.3
7	13.2	15.1	14.0

Pruebe si existe diferencia entre las valuaciones, con un nivel de significación de 0.01.

4. Una empresa de taxis evalúa los tiempos que se lleva recorrer 4 rutas alternativas entre el aeropuerto y el Centro de Convenciones de una ciudad grande y, para ello, tomó los tiempos de recorrido en determinada hora y para los 5 días hábiles de la semana, con el mismo conductor, y obtuvo los siguientes resultados, en minutos:

	Ruta A	Ruta B	Ruta C	Ruta D
Lunes	42	42	43	43
Martes	45	44	47	45
Miércoles	44	45	48	44
Jueves	48	43	51	46
Viernes	49	47	50	47

Pruebe si existe diferencia entre los tiempos de recorrido de las 4 rutas, con un nivel de significación de 0.01.

5. Para evaluar el rendimiento de 5 operadores experimentados con 5 diferentes modelos de máquinas empacadoras de modelos distintos, se les pidió a los operadores hacer pruebas con los 5 modelos de máquinas y se observó que el número de cajas empacadas por turno fue:

	Máquina				
Operador	**I**	**II**	**III**	**IV**	**V**
A	325	293	296	299	329
B	298	301	305	323	325
C	302	305	313	334	349
D	313	322	335	342	359
E	305	307	310	315	308

Pruebe esa hipótesis con un nivel de significación de 2.5 por ciento.

17.8 Coeficiente de correlación por rangos de Spearman

Este coeficiente de correlación por rangos es el equivalente no paramétrico del coeficiente de correlación que se analizó en la sección 13.10 y que se calcula para variables que están cuando menos en escala de intervalo y se utiliza, de igual manera, para medir la relación que existe entre 2 variables y mide tanto la intensidad como el sentido de la relación. Este coeficiente de correlación por rangos se calcula para variables que están en escala ordinal y asume los mismos valores que su contraparte paramétrica; es decir, asume valores entre –1 y 1, en donde un valor de –1 significa que existe una correlación negativa perfecta entre las 2 variables, un valor de 1 señala que existe una correlación positiva perfecta y valores de 0 o cercanos a 0 son señal de que no existe o existe poca relación entre las 2 variables.

El procedimiento para calcular este coeficiente de correlación por rangos es similar a las otras pruebas que ya se revisaron y que se basan en rangos: se asignan rangos a las observaciones de las 2 variables y los empates se resuelven asignando el promedio de los rangos que les corresponderían en caso de no haber empates. Después de que se obtienen los rangos, se calculan las diferencias de éstos para cada par de observaciones de las 2 variables. El estadístico de prueba, el coeficiente de correlación de Spearman, es:

$$r_r = 1 - \frac{6\sum d^2}{n\left(n^2 - 1\right)} \tag{17.17}$$

Cuando la muestra tiene cuando menos 10 observaciones se puede utilizar la t de la distribución t de Student como estadístico teórico de prueba:

$$t = \frac{r_r}{\sqrt{\left(\frac{1 - r_r^2}{n - 2}\right)}} \tag{17.18}$$

Con grados de libertad $n - 2$.

Ejemplo 17.16

Se toma una muestra aleatoria de 11 operarios fabriles y se anotan las calificaciones de desempeño que les otorgan 2 supervisores, con los resultados que se anotan en la tabla 17.8.

Tabla 17.18 Calificaciones de 11 operarios fabriles, dadas por 2 supervisores

Operario	Calificación del supervisor A	Calificación del supervisor B
1	81	78
2	83	83
3	90	92
4	98	72
5	78	74
6	74	80
7	85	84
8	90	79
9	95	93
10	91	94
11	92	95

Calcule el coeficiente de correlación por rangos de Spearman y pruebe, con un nivel de significación de 5% si es estadísticamente significativo.

Solución: Las hipótesis:

H_0: $\rho_r = 0$
H_0: $\rho_r \neq 0$

En la tabla 17.19 se resumen los cálculos.

Tabla 17.19 Cálculos para el ejemplo 17.16

Operario	Calificación A	Calificación B	Rango A	Rango B	d	d^2
1	81	78	3	2	1	1
2	83	83	4	5	–1	1
3	90	92	6.5	7.5	–1	1
4	98	92	11	7.5	3.5	12.25
5	78	74	2	1	1	1
6	74	80	1	4	–3	9
7	85	84	5	6	–1	1
8	90	79	6.5	3	3.5	12.25
9	95	93	10	9	1	1
10	91	94	8	10	–2	4
11	92	95	9	11	–2	4
					Suma	47.5

Sustituyendo ahora en la ecuación 17.17:

$$r_r = 1 - \frac{6\sum d^2}{n\left(n^2 - 1\right)} = 1 - \frac{6(47.5)}{11(121 - 1)} = 1 - \frac{285}{1\,320} = 0.7841$$

El valor calculado del estadístico de prueba:

$$= \frac{r_r}{\sqrt{\frac{1 - r_r^2}{n - 2}}} = \frac{0.7841}{\sqrt{\frac{1 - (0.7841)^2}{11 - 2}}} = \frac{0.7841}{\sqrt{\frac{0.3852}{9}}} = \frac{0.7841}{0.2069} = 3.7900$$

El valor crítico del estadístico de prueba es $t = 2.262$, ya que $P(-2.262 \leq t \leq 2.262 \mid gl = 9) = 0.05$.

Como el valor calculado del estadístico de prueba, 3.7900, es mayor que 2.262, se rechaza la hipótesis nula y se concluye que el coeficiente de correlación por rangos de Spearman no es igual a 0 y, por lo tanto, se concluye también que sí existe relación entre las calificaciones de los 2 supervisores.

ejercicios 17.8 Coeficiente de correlación por rangos de Spearman

1. Se determinó el número de habitantes por kilómetro cuadrado y la calidad del aire (medida como 1: muy baja, 2: baja, 3: normal y 4: buena) en 10 ciudades de la República Mexicana y se obtuvieron los siguientes resultados:

Ciudad	Habitantes (miles)	Calidad del aire
1	2 782	1
2	869	2
3	924	1
4	573	3
5	668	4
6	321	4
7	524	3
8	399	4
9	464	2
10	721	2

¿Se puede afirmar que a mejor calidad del aire menor número de habitantes por km^2? Realice la prueba con un nivel de significación de $\alpha = 0.05$. Si existe la relación, describa de qué tipo es.

2. En un estudio de tipo genético se desea determinar la posible asociación entre la estatura del padre y la del hijo mayor. Se midieron ambas variables en centímetros y se obtuvieron los siguientes resultados:

187.96	185.42
180.34	180.34
182.88	193.04
180.34	177.80
175.26	182.88
165.10	172.72
182.88	175.26
172.72	175.26
177.80	177.80
170.18	175.26

Pruebe, con nivel de significación $\alpha = 0.05$, si existe relación entre estas 2 variables. Si existe la relación, describa de qué tipo es.

3. Del reporte de calificaciones del curso propedéutico de la generación anterior, se tomaron las calificaciones de 17 estudiantes, tanto a la mitad del curso, como al final y se obtuvieron los siguientes resultados:

	Calificación	
Estudiante	A mitad del curso	Al final del curso
1	6.9	8.6
2	1.4	4.6
3	6.8	7.2
4	6.1	7.6
5	2.5	3.5
6	9.8	9.0
7	5.8	6.3
8	5.7	5.0
9	5.7	6.3
10	5.4	7.9
11	5.7	6.6
12	7.1	6.2
13	6.0	6.4
14	7.3	8.5
15	8.0	7.7
16	7.5	7.0
17	7.1	8.0

Pruebe si las 2 calificaciones se relacionan con un nivel de significación de $\alpha = 0.05$.

4. La revista *Expansión* publica cada año la lista "Las 500 empresas más importantes de México". En la tabla siguiente se anotan algunos datos de las 10 primeras que componen la lista, en orden descendente de sus ventas. Pruebe si existe relación entre su posición en el listado y el número de empleados, utilizando un nivel de significación de 1 por ciento.

	Empresa	Ubicación	Sector	País	Ventas (mdp)	Empleados
1	Petróleos Mexicanos	DF	Minería, petróleo y gas	MX	1 282 064.30	147 672
2	América Móvil	DF	Medios y telecomunicaciones	MX	607 855.70	150 618

	Empresa	Ubicación	Sector	País	Ventas (mdp)	Empleados
3	Walmart de México	DF	Comercio minorista	EU	335 857.40	219 767
4	Comisión Federal de Electricidad	DF	Electricidad, agua y gas	MX	254 417.30	93 254
5	Cemex	NL	Cemento, cerámica y vidrio	MX	178 260.00	46 523
6	Fomento Económico Mexicano	NL	Bienes de consumo	MX	169 701.80	108 572
7	General Motors de México	DF	Automotriz y autopartes	EU	158 692.00	12 000
8	Grupo Alfa	NL	Holding	MX	136 395.00	56 332
9	Grupo Financiero BBVA-Bancomer	DF	Servicios financieros y seguros	ESP	121 910.00	34 189
10	Ford Motor Company	DF	Automotriz y autopartes	EU	121 000.00	7 677

Fuente: CNN Expansión, "Las 500 empresas más importantes de México", disponible en: http://www.cnnexpansion.com/rankings/2011/las-500-empresas-mas-importantes-de-mexico-2011/ranking.php, consultado el 12 de agosto de 2012.

5. Un despacho de reclutamiento de personal aplica a los solicitantes 2 evaluaciones antes de proponerlos para empleo en empresas. Una primera evaluación se basa en el potencial que el despacho les asigna con base en una evaluación curricular y la otra es con base en una batería de exámenes diversos que incluyen habilidades mentales, de comunicación, y otras. Las calificaciones que obtuvieron 12 solicitantes son:

A	8	4
B	10	4
C	9	4
D	4	3
E	12	6
F	11	9
G	11	9
H	7	6
I	8	6
J	13	9
K	10	5
L	12	9
M	8	4

Pruebe, con un nivel de significación de 0.01, si existe relación entre las 2 evaluaciones.

17.9 Resumen

Se revisaron las principales pruebas de hipótesis no paramétricas, las cuales se distinguen de las paramétricas porque no poseen cuando menos una de las características que suelen tener las pruebas paramétricas, es decir:

- No se hace la prueba sobre un parámetro.
- La escala de medición de la o las variables es nominal u ordinal.
- No se requieren suposiciones sobre la forma de la distribución de la variable en la población.

En la sección 17.1 se incluye un cuadro en donde se listan las pruebas no paramétricas que se estudian en este texto, incluyendo las de este capítulo y las que se estudiaron en el capítulo 11 y que se llevan a cabo utilizando la χ^2, ji cuadrada, como estadístico de prueba.

17.10 Fórmulas del capítulo

17.2 Prueba de rachas para aleatoriedad de Wald-Wolfowitz

Media, error estándar y estadístico de prueba para la y prueba de rachas para aleatoriedad:

$$_r = \frac{2n_1n_2}{n_1+n_2}+1 \qquad \textbf{(17.1)}$$

$$\sigma_r = \sqrt{\frac{2n_1n_2(2n_1n_2-n_1-n_2)}{(n_1+n_2)^2(n_1+n_2-1)}} \qquad \textbf{(17.2)}$$

Por lo que se puede utilizar la z normal como estadístico de prueba:

$$z = \frac{r - \mu_r}{\sigma_r} \tag{17.3}$$

Fórmula para el cálculo de la probabilidad de obtener determinado número de rachas, cuando son pares, con $r = 2k$:

$$P(r) = \frac{2\begin{pmatrix} n_1 - 1 \\ k - 1 \end{pmatrix}\begin{pmatrix} n_2 - 1 \\ k - 1 \end{pmatrix}}{\begin{pmatrix} n_1 + n_2 \\ n_1 \end{pmatrix}} \tag{17.4}$$

Fórmula para el cálculo de la probabilidad de obtener determinado número de rachas, cuando son nones, con $r = 2k + 1$:

$$P(r) = \frac{\begin{pmatrix} n_1 - 1 \\ k \end{pmatrix}\begin{pmatrix} n_2 - 1 \\ k - 1 \end{pmatrix} + \begin{pmatrix} n_1 - 1 \\ k - 1 \end{pmatrix}\begin{pmatrix} n_2 - 1 \\ k \end{pmatrix}}{\begin{pmatrix} n_1 + n_2 \\ n_1 \end{pmatrix}} \tag{17.5}$$

17.3 Prueba de los signos

El estadístico de prueba para la prueba de los signos, con muestras grandes $n \geq 10$:

$$z = \frac{(p \pm 0.005) - \pi}{\sqrt{\frac{(0.25)}{n}}} \tag{17.6}$$

17.4 Prueba de rangos con signo de Wilcoxon

Media, error estándar y estadístico de prueba para la prueba de rangos con signo de Wilcoxon, con muestras grandes:

$$z = \frac{T - \mu_T}{s_T} \tag{17.7}$$

en donde

$$\mu_T = \frac{n(n+1)}{4} \tag{17.8}$$

y

$$\sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{24}} \tag{17.9}$$

17.5 Prueba U de Mann-Whitney para 2 muestras independientes

El estadístico U de Mann-Whitney para una prueba de un extremo:

$$U_1 = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1 \tag{17.10}$$

El estadístico U de Mann-Whitney para una prueba de 2 extremos:

$$U_2 = n_1 n_2 + \frac{n_2(n_2 + 1)}{2} - R_2 \tag{17.11}$$

Media, error estándar y estadístico de prueba para la prueba U de Mann-Whitney, con muestras grandes:

$$z = \frac{U_1 - \mu_{U_1}}{\sigma_{U_1}} \tag{17.12}$$

en donde

$$_{U_1} = \frac{n_1 n_2}{2} \tag{17.13}$$

y

$$\sigma_{U_1} = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} \tag{17.14}$$

17.6 Prueba de suma de rangos de Kruskal-Wallis para más de 2 medias

El estadístico de prueba para la prueba de la suma de rangos de Kruskal-Wallis:

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_1^2}{n_i} - 3(n+1) \tag{17.15}$$

17.7 Prueba de Friedman para diseños en bloques aleatorizados

El estadístico de prueba para la prueba de suma de rangos de Friedman para análisis de varianza 2 factores:

$$F_r = \frac{12}{nk(k+1)} \sum_{j=1}^{k} R_j^2 - 3n(k+1) \tag{17.16}$$

17.8 Coeficiente de correlación por rangos de Spearman

El coeficiente de correlación de Spearman:

$$r_r = 1 - \frac{6\sum d^2}{n(n^2 - 1)} \tag{17.17}$$

El estadístico de prueba para pruebas de hipótesis sobre el coeficiente de correlación de Spearman:

$$t = \frac{r_r}{\sqrt{\frac{(1 - r_r^2)}{n - 2}}} \tag{17.18}$$

17.11 Ejercicios adicionales

17.2 Pruebas de rachas para aleatoriedad de Wald-Wolfowitz

1. Se lanzó una moneda 36 veces y se obtuvieron los siguientes resultados (S = una cara de la moneda, A = la otra cara).

A	A	S	S
S	S	A	A
S	S	S	S
S	S	S	A
S	A	S	A
A	A	A	S
A	S	A	S
A	S	S	S
S	S	S	A

Utilice la prueba de rachas para aleatoriedad de Wald-Wolfowitz para probar si se puede afirmar que la moneda no tiene truco, utilizando un nivel de significación de 5 por ciento.

2. En la tabla siguiente se muestran los montos de 40 facturas elegidas al azar, utilice la prueba de rachas para aleatoriedad para probar si se puede considerar que la secuencia de los montos es aleatoria, utilizando un nivel de significación de 1% y considerando que la mediana de esas cantidades es 96.

93	100	35	222
51	55	119	95
45	97	30	211
190	66	161	44
120	72	152	78
127	138	100	87
254	85	189	63
58	32	93	300
165	142	59	33
121	72	365	159

3. Para evaluar si las diferencias en el contenido de los paquetes de té que una empresa vende, se tomó una muestra aleatoria de 20 y se obtuvieron los siguientes resultados:

37.59
36.75
36.12
36.33
34.65
35.28
35.07
36.12
36.54
36.96
36.75
37.38
35.28
34.65
34.86
37.17
36.96
37.17
37.38
36.12

Con un nivel de significación de $\alpha = 0.01$, ¿se puede decir que la distribución de los pesos de los paquetes es aleatoria alrededor de la mediana general de 36.5 gramos?

17.3 Prueba de los signos
17.3.2 Prueba del signo para una muestra pequeña

4. En una línea de ensamble se prueba una nueva técnica que se supone aumenta el número de piezas ensambladas que era, antes de la aplicación de la nueva técnica, de 90 unidades por turno. Con los datos obtenidos utilizándola y que aparecen en seguida, pruebe, con un nivel de significación, si la nueva técnica funciona.

Turno	Piezas ensambladas
1	85
2	95
3	102
4	90
5	104
6	100
7	101
8	86

5. Los resultados que se mostrarán en seguida corresponden a los resultados que obtuvieron 9 personas que se sometieron a una dieta que se anuncia por televisión:

Persona	Resultado
Dondé	Subió
Vázquez	Bajó
López	Subió
Jasso	Bajó
Álvarez	Sin cambio
Jorrín	Bajó
Franco	Bajó
Smith	Bajó
Simón	Sin cambio

a) Plantee la hipótesis pertinente y pruébela con un nivel de significación de 0.01.
b) Interprete el resultado de la prueba.

6. Para evaluar la efectividad de un curso de ortografía impartido a los estudiantes de un grupo de secretarias, se evaluó la calidad de su ortografía antes y después del curso y, a partir de una muestra de 10 de ellas, se obtuvieron los siguientes resultados:

Secretaria	La calidad de la ortografía
A	Mejoró
B	Igual
C	Subió
D	Bajó
E	Mejoró
F	Mejoró
G	Mejoró
H	Bajó
I	Mejoró
J	Igual

Con un nivel de significación de 0.05, ¿se puede afirmar que el curso de ortografía fue útil?

17.3.3 Prueba del signo para una muestra grande (aproximación normal)

7. Las cantidades vendidas de cierto producto en los 12 meses anteriores fueron:

88
198
99
132
110
154
176
77
154
121
110
220

Utilice la prueba de los signos para probar la hipótesis de que la mediana de los productos vendidos por mes no es superior a 110 por mes, con un nivel de significación de 0.01.

8. Se supone que la edad mediana de un grupo grande de adultos es de 43 años. Pruebe esta hipótesis, con un nivel de significación de 1%, utilizando las edades de una muestra aleatoria de las siguientes 15 personas:

46.4
48.3
51.9
38.8
46.5
45.6
52.1
41
54.2
44.9
52.3
43.6
48.7
42.2
44.9

9. Los datos siguientes son el número de defectos que se encontraron en una muestra de 25 rollos de tela. Se asume que el proceso de producción está bajo control si la mediana de los defectos en los rollos fabricados no excede de 5. ¿Se puede afirmar que el proceso de producción de donde procede esta muestra de rollos está bajo control? Utilice un nivel de significación de 0.05.

3	4	13	7	1
1	4	14	8	11
2	5	12	6	10
5	6	13	9	10
3	2	12	7	11

17.3.4 Prueba del signo para 2 muestras apareadas pequeñas

10. Se pidió a un grupo de 9 amas de casa que evaluaran la eficiencia de 2 jabones líquidos para trastes y que los calificaran en una escala de 0 a 10, con los siguientes resultados:

Ama de casa	Jabón 1	Jabón 2
1	8	7
2	7	9
3	6	7
4	8	7
5	6	8
6	6	8
7	6	7
8	7	9
9	6	8

Pruebe, a un nivel de significación de 0.05, la hipótesis nula de que no existe diferencia entre las calificaciones asignadas por las amas de casa a los 2 jabones.

11. Las calificaciones siguientes se obtuvieron al pedirle a 10 concursantes que evaluaran la dificultad para resolver un acertijo con y sin pista inicial:

Concursante	Con pista inicial	Sin pista inicial
1	86	88
2	92	84
3	94	98
4	82	96
5	80	88
6	84	84
7	86	95
8	94	84
9	81	90
10	82	90

Demuestre, con la prueba de los signos y con un nivel de significación de 0.05, si existe diferencia en la dificultad para resolver el acertijo con y sin pista inicial.

12. Se implantó un curso de capacitación avanzada para el manejo de paquetería de computación entre los empleados encargados de la formación editorial de materiales para su publicación. Posteriormente se compararon datos sobre la cantidad de cuartillas que terminaba cada empleado, en turnos de 4 horas, antes de la capacitación y después de ella y se encontró la información que se muestra en la tabla siguiente.

	Núm. de cuartillas	
Empleado	Antes	Después
A	19	20
B	23	25
C	27	24
D	17	27
E	12	30
F	18	18
G	12	24
H	22	21
I	19	22
J	26	32

Demuestre, con la prueba de los signos y con un nivel de significación de 0.01, si existe diferencia en la productividad de los empleados, antes y después de la capacitación.

17.3.5 Prueba del signo para 2 muestras apareadas grandes (aproximación normal)

13. Para evaluar la efectividad de 2 tipos de promoción en tiendas de conveniencia, se tomó una muestra aleatoria de 15 tiendas en las cuales se ensayó durante una semana el uso de toda clase de publicidad impresa y en la semana siguiente se utilizaron botargas, y se midieron las ventas semanales, mismas que se muestran en la tabla siguiente.

Tienda	Publicidad impresa	Botargas
1	413 113	415 115
2	429 229	405 205
3	402 802	485 785
4	363 063	486 987
5	481 781	438 038
6	392 893	484 284
7	484 484	352 953
8	373 473	430 430
9	378 278	467 667
10	423 323	472 572
11	398 098	477 477
12	378 178	448 848
13	360 560	439 339
14	396 296	456 456
15	492 292	385 185

Pruebe, a un nivel de significación de 0.01, la hipótesis nula de que no existe diferencia entre las ventas semanales durante las 2 semanas en las que se utilizaron las 2 formas de promoción.

14. En un estudio para probar los efectos de la música ambiental sobre el desempeño de trabajadores fabriles, se midió el número promedio de productos que cada uno de ellos terminaba cada día, durante los 30 días previos a la instalación de música ambiental, y el mismo promedio durante otros 30 días durante los cuales sí había música ambiental. Los resultados fueron los siguientes:

Trabajador	Sin música	Con música
1	55	64
2	45	50
3	57	63
4	50	48
5	46	53
6	50	64
7	37	45
8	50	56
9	35	45
10	53	59
11	62	60
12	55	57

¿Se puede afirmar, con un nivel de significación de 0.01, que no existe diferencia entre los promedios de productos terminados por trabajador, en las 2 condiciones?

15. Para evaluar el gusto de los consumidores por 2 sabores para una golosina, a una muestra aleatoria de 20 personas se les pidió evaluar 2 golosinas con cada uno de los 2 sabores y que manifestaran cuál les gustaba más.

Persona	Sabor A	Sabor B
1	X	
2	X	
3		X

(*continúa*)

(*continuación*)

Persona	Sabor A	Sabor B
4		X
5	Sin preferencia	Sin preferencia
6	X	
7	X	
8	X	
9	X	
10	X	
11		X
12	X	
13	X	
14	Sin preferencia	Sin preferencia
15	X	
16	X	
17	X	
18	X	
19	X	
20		X

Demuestre, a un nivel de significación de 0.05, la hipótesis nula de que no existe diferencia entre las preferencias que los consumidores manifestaron sobre los 2 sabores.

17.4 Prueba de rangos con signo de Wilcoxon
17.4.2 Prueba de rangos con signo de Wilcoxon para una muestra pequeña

16. Resuelva el ejercicio 4 utilizando esta prueba de Wilcoxon; compare los resultados obtenidos en ambos casos y comente.

17. Resuelva el ejercicio 5 utilizando esta prueba de Wilcoxon; compare los resultados obtenidos en ambos casos y comente.

18. Resuelva el ejercicio 6 utilizando esta prueba de Wilcoxon; compare los resultados obtenidos en ambos casos y comente.

17.4.4 Prueba de rangos con signo de Wilcoxon para una muestra grande (aproximación normal)

19. Se fabrican artículos mediante un proceso tradicional que logra una producción mediana de 80 artículos por turno. Los datos de la tabla que se reproduce en seguida muestran el número de artículos fabricados en 15 turnos muestreados al azar y en los cuales se utilizó un nuevo programa de "trabajo por equipos" propuesto por los trabajadores y que, según afirman, permite aumentar el nivel de producción.

Turno	Producción
1	77
2	85
3	90
4	78
5	92
6	92
7	89
8	78
9	89
10	80
11	94
12	81
13	76
14	85
15	80

Pruebe si realmente ha aumentado la producción con un nivel de significación de $\alpha = 0.05$.

20. Resuelva el ejercicio 7 utilizando esta prueba de Wilcoxon; compare los resultados obtenidos en ambos casos y comente.

21. Resuelva el ejercicio 8 utilizando esta prueba de Wilcoxon; compare los resultados obtenidos en ambos casos y comente.

22. Resuelva el ejercicio 9 utilizando esta prueba de Wilcoxon; compare los resultados obtenidos en ambos casos y comente.

17.4.5 Prueba de rangos con signo de Wilcoxon para 2 muestras apareadas pequeñas

23. Resuelva el ejercicio 10 utilizando esta prueba de Wilcoxon; compare los resultados obtenidos en ambos casos y comente.

24. Resuelva el ejercicio 11 utilizando esta prueba de Wilcoxon; compare los resultados obtenidos en ambos casos y comente.

25. Resuelva el ejercicio 12 utilizando esta prueba de Wilcoxon; compare los resultados obtenidos en ambos casos y comente.

17.4.6 Prueba de rangos con signo de Wilcoxon para 2 muestras apareadas grandes (aproximación normal)

26. Resuelva el ejercicio 13 utilizando esta prueba de Wilcoxon; compare los resultados obtenidos en ambos casos y comente.

27. Resuelva el ejercicio 14 utilizando esta prueba de Wilcoxon; compare los resultados obtenidos en ambos casos y comente.

28. Resuelva el ejercicio 15 utilizando esta prueba de Wilcoxon; compare los resultados obtenidos en ambos casos y comente.

17.5 Prueba de Mann-Whitney para 2 muestras independientes

29. Se evaluaron las habilidades de mujeres y hombres en diversas tareas en 2 muestras aleatorias de 20 elementos para cada sexo y se obtuvieron las siguientes calificaciones globales:

Mujeres	Hombres
93.50	100.00
89.00	98.00
95.00	91.20
94.00	94.75
94.50	95.00

Mujeres	Hombres
92.00	94.10
91.00	93.75
87.50	96.10
87.00	92.50
85.00	92.75
87.40	90.00
90.20	96.00
87.50	93.50
94.25	92.25
86.00	91.30

Utilice la prueba de Mann-Whitney para probar si existe una diferencia real entre el desempeño de los 2 sexos, a un nivel de significación de $\alpha = 0.05$.

30. Después de que 2 grupos de trabajadores terminaron un curso de capacitación utilizando 2 métodos distintos, se aplicó a los trabajadores capacitados el mismo examen y se obtuvieron los siguientes resultados:

Capacitación tipo I	Capacitación tipo II
6.93	8.51
8.91	7.72
8.12	8.91
6.34	8.12
8.51	6.44
7.62	8.61
8.32	7.92
7.82	8.71
8.12	9.41
8.81	8.42
7.23	7.52
8.02	9.31
8.22	6.53

Con un nivel de significación de 0.01, pruebe la hipótesis de que en ambos grupos se obtuvo la misma calificación mediana.

31. En una estación de verificación vehicular de emisiones contaminantes se prueba un equipo nuevo de verificación y se tomaron muestras del tiempo que lleva hacer todo el procedimiento con el equipo antiguo y con el nuevo, y se obtuvieron los siguientes tiempos:

15	8
19	27
32	30
6	46
21	39
4	18
23	10
35	24
16	29
24	37
20	21
33	55
18	28
9	22
39	14

Pruebe, con un nivel de significación de 0.05, si ambas muestras tienen la misma mediana y concluya según proceda, en términos de la eficiencia del equipo nuevo.

17.6 Prueba de Kruskal-Wallis para más de 2 medias

32. Se realizaron pruebas para evaluar si 4 diferentes métodos de producción eran igualmente eficientes. Se muestran en seguida los números de artículos defectuosos producidos en sendas corridas de 10 000 artículos cada una, fabricados con los distintos métodos.

Método			
I	II	III	IV
78	97	87	74
86	89	80	75
85	96	79	73
84	96	78	76
88	97	84	74
86	94	84	71
83	90	84	69
	96	82	78
			73

Pruebe la hipótesis de que los 4 métodos producen el mismo número de artículos defectuosos por corrida, con un nivel de significación de 5 por ciento.

33. Se probaron 4 presentaciones de determinado producto, solicitando a 6 evaluadores de una agencia de publicidad que los calificaran en una escala de 0 a 100 y se obtuvieron los siguientes resultados:

Presentaciones			
I	II	III	IV
71	84	93	90
87	88	93	92
69	85	87	91
86	81	86	94
70	87	88	93
75	85	89	91

Pruebe la hipótesis de que no existen diferencias entre las calificaciones otorgadas por los diferentes evaluadores, con un nivel de significación de 0.05.

34. Una empresa que provee servicios de televisión por cable trata de determinar si existen diferencias entre los patrones de pago de sus clientes morosos. Para ello, tome una muestra aleatoria de las facturas de 10 clientes morosos en cada una de 4 ciudades y determine el número de días de retraso que tiene cada una, con los resultados que se muestran a continuación:

Ciudad 1	Ciudad 2	Ciudad 3	Ciudad 4
21	7	5	1
11	7	7	7
15	8	13	8
20	5	7	13
19	7	10	6
12	7	7	2
11	3	6	6
17	7	11	6
11	8	13	3
17	4	10	7

Con un nivel de significación de 0.05, ¿existen diferencias entre los tiempos de demora de los clientes morosos en las 4 ciudades?

17.7 Prueba de Friedman para diseños en bloques aleatorizados

35. Se pidió a una muestra de 12 personas elegidas al azar que calificaran 5 destinos de playa en términos de diversos aspectos y se obtuvieron las siguientes calificaciones:

	Destino				
Evaluador	I	II	III	IV	V
1	5	6	4	7	8
2	6	5	7	8	4
3	6.5	6.5	8	5	4
4	5	4	6	7	8
5	6.5	6.5	4	8	5
6	6	7	4	8	5
7	8	4	7	6	5
8	8	6	4	5	7
9	5	8	4	7	6
10	5	8	6	4	7
11	6	8	7	5	4
12	7	5	4	6	8

Pruebe, con un nivel de significación de 0.01, si existen diferencias entre las calificaciones de los 5 destinos.

36. Para evaluar la eficiencia en el blanqueo de pulpa de papel, se ensayaron 3 combinaciones de agentes blanqueadores en 5 tipos de pulpa de papel y se tomaron medidas del nivel de blanqueo:

Tipos de pulpa de papel	Combinaciones de agentes blanqueadores		
	A	B	C
I	7.79	8.53	10.00
II	8.21	9.05	10.00
III	8.00	9.47	9.47
IV	8.63	9.79	9.16
V	8.11	7.68	9.79

Pruebe, con un nivel de significación de 0.05, si existen diferencias en la eficacia entre las 3 combinaciones de agentes blanqueadores.

37. En una empresa de servicio que trabaja las 24 horas se decidió ofrecer diversos conjuntos de prestaciones para los trabajadores de los diferentes puestos de los 3 turnos con el propósito de hacer que los 3 turnos de trabajo les resultaran igualmente atractivos.

Tiempo después de la implantación del plan se realizó una encuesta al azar entre trabajadores de los 6 tipos de puestos que laboraron en los 3 turnos para evaluar qué tan satisfechos estaban con éstos. Los resultados que se obtuvieron son los que se muestran en seguida:

Puesto	Matutino	Vespertino	Nocturno
1	67	81	81
2	68	64	84
3	69	77	90
4	65	72	86
5	73	84	78
6	72	81	84

Con un nivel de significación de $\alpha = 0.05$, pruebe si existen diferencias entre las calificaciones otorgadas a los 3 turnos.

17.8 Coeficiente de correlación de rangos de Spearman

38. En la tabla siguiente se presentan los datos obtenidos de una muestra aleatoria de 9 empleados de una compañía, para los cuales se obtuvo, por un lado, una medición del tiempo de capacitación y experiencia y, por el otro lado, de su rendimiento semanal.

Calcule el coeficiente de correlación por rangos de Spearman y pruebe, con un nivel de significación de 1%, si es estadísticamente significativo.

	Calificaciones	
Empleado	Capacitación y experiencia	Rendimiento semanal
1	95	83
2	81	87
3	71	56
4	86	95
5	67	53
6	68	65
7	74	80

	Calificaciones	
Empleado	Capacitación y experiencia	Rendimiento semanal
8	58	68
9	71	77

39. Se formaron 8 grupos de evaluación con el propósito de calificar 2 anuncios para radio que los creativos de una agencia de publicidad proponen para promover un producto.

Grupo	Anuncio A	Anuncio B
1	7	7
2	5	10
3	9	9
4	8	3
5	3	8
6	2	2
7	4	6
8	6	10

Pruebe si se correlacionan las calificaciones de los 2 anuncios con un nivel de significación de 0.05.

40. Se pidió a una muestra de parejas del mismo grupo de edad que calificaran 12 programas de televisión, con los resultados que se muestran a continuación:

Programa	Mujeres	Hombres
A	96	95
B	94	96
C	93	92
D	98	93
E	88	89
F	92	94
G	95	97
H	97	91
I	87	98
J	86	90
K	99	99
L	91	87

Con un nivel de significación de 0.05, pruebe si se correlacionan las calificaciones de hombres y mujeres.

Apéndices

Apéndice 1

Tabla de probabilidades nominales

		Probabilidad										
		0.01	**0.05**	**0.10**	**0.15**	**0.20**	**0.25**	**0.30**	**0.35**	**0.40**	**0.45**	**0.50**
1	0	0.9900	0.9500	0.9000	0.8500	0.8000	0.7500	0.7000	0.6500	0.6000	0.5500	0.5000
1	1	0.0100	0.0500	0.1000	0.1500	0.2000	0.2500	0.3000	0.3500	0.4000	0.4500	0.5000
2	0	0.9801	0.9025	0.8100	0.7225	0.6400	0.5625	0.4900	0.4225	0.3600	0.3025	0.2500
2	1	0.0198	0.0950	0.1800	0.2550	0.3200	0.3750	0.4200	0.4550	0.4800	0.4950	0.5000
2	2	0.0001	0.0025	0.0100	0.0225	0.0400	0.0625	0.0900	0.1225	0.1600	0.2025	0.2500
3	0	0.9703	0.8574	0.7290	0.6141	0.5120	0.4219	0.3430	0.2746	0.2160	0.1664	0.1250
3	1	0.0294	0.1354	0.2430	0.3251	0.3840	0.4219	0.4410	0.4436	0.4320	0.4084	0.3750
3	2	0.0003	0.0071	0.0270	0.0574	0.0960	0.1406	0.1890	0.2389	0.2880	0.3341	0.3750
3	3	0.0000	0.0001	0.0010	0.0034	0.0080	0.0156	0.0270	0.0429	0.0640	0.0911	0.1250
4	0	0.9606	0.8145	0.6561	0.5220	0.4096	0.3164	0.2401	0.1785	0.1296	0.0915	0.0625
4	1	0.0388	0.1715	0.2916	0.3685	0.4096	0.4219	0.4116	0.3845	0.3456	0.2995	0.2500
4	2	0.0006	0.0135	0.0486	0.0975	0.1536	0.2109	0.2646	0.3105	0.3456	0.3675	0.3750
4	3	0.0000	0.0005	0.0036	0.0115	0.0256	0.0469	0.0756	0.1115	0.1536	0.2005	0.2500
4	4	0.0000	0.0000	0.0001	0.0005	0.0016	0.0039	0.0081	0.0150	0.0256	0.0410	0.0625
5	0	0.9510	0.7738	0.5905	0.4437	0.3277	0.2373	0.1681	0.1160	0.0778	0.0503	0.0313
5	1	0.0480	0.2036	0.3281	0.3915	0.4096	0.3955	0.3602	0.3124	0.2592	0.2059	0.1563
5	2	0.0010	0.0214	0.0729	0.1382	0.2048	0.2637	0.3087	0.3364	0.3456	0.3369	0.3125
5	3	0.0000	0.0011	0.0081	0.0244	0.0512	0.0879	0.1323	0.1811	0.2304	0.2757	0.3125
5	4	0.0000	0.0000	0.0005	0.0022	0.0064	0.0146	0.0284	0.0488	0.0768	0.1128	0.1563
5	5	0.0000	0.0000	0.0000	0.0001	0.0003	0.0010	0.0024	0.0053	0.0102	0.0185	0.0313
6	0	0.9415	0.7351	0.5314	0.3771	0.2621	0.1780	0.1176	0.0754	0.0467	0.0277	0.0156
6	1	0.0571	0.2321	0.3543	0.3993	0.3932	0.3560	0.3025	0.2437	0.1866	0.1359	0.0938
6	2	0.0014	0.0305	0.0984	0.1762	0.2458	0.2966	0.3241	0.3280	0.3110	0.2780	0.2344
6	3	0.0000	0.0021	0.0146	0.0415	0.0819	0.1318	0.1852	0.2355	0.2765	0.3032	0.3125
6	4	0.0000	0.0001	0.0012	0.0055	0.0154	0.0330	0.0595	0.0951	0.1382	0.1861	0.2344
6	5	0.0000	0.0000	0.0001	0.0004	0.0015	0.0044	0.0102	0.0205	0.0369	0.0609	0.0938
6	6	0.0000	0.0000	0.0000	0.0000	0.0001	0.0002	0.0007	0.0018	0.0041	0.0083	0.0156
7	0	0.9321	0.6983	0.4783	0.3206	0.2097	0.1335	0.0824	0.0490	0.0280	0.0152	0.0078

(*continúa*)

(*continuación*)

		Probabilidad										
		0.01	**0.05**	**0.10**	**0.15**	**0.20**	**0.25**	**0.30**	**0.35**	**0.40**	**0.45**	**0.50**
7	1	0.0659	0.2573	0.3720	0.3960	0.3670	0.3115	0.2471	0.1848	0.1306	0.0872	0.0547
7	2	0.0020	0.0406	0.1240	0.2097	0.2753	0.3115	0.3177	0.2985	0.2613	0.2140	0.1641
7	3	0.0000	0.0036	0.0230	0.0617	0.1147	0.1730	0.2269	0.2679	0.2903	0.2918	0.2734
7	4	0.0000	0.0002	0.0026	0.0109	0.0287	0.0577	0.0972	0.1442	0.1935	0.2388	0.2734
7	5	0.0000	0.0000	0.0002	0.0012	0.0043	0.0115	0.0250	0.0466	0.0774	0.1172	0.1641
7	6	0.0000	0.0000	0.0000	0.0001	0.0004	0.0013	0.0036	0.0084	0.0172	0.0320	0.0547
7	7	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0002	0.0006	0.0016	0.0037	0.0078
8	0	0.9227	0.6634	0.4305	0.2725	0.1678	0.1001	0.0576	0.0319	0.0168	0.0084	0.0039
8	1	0.0746	0.2793	0.3826	0.3847	0.3355	0.2670	0.1977	0.1373	0.0896	0.0548	0.0313
8	2	0.0026	0.0515	0.1488	0.2376	0.2936	0.3115	0.2965	0.2587	0.2090	0.1569	0.1094
8	3	0.0001	0.0054	0.0331	0.0839	0.1468	0.2076	0.2541	0.2786	0.2787	0.2568	0.2188
8	4	0.0000	0.0004	0.0046	0.0185	0.0459	0.0865	0.1361	0.1875	0.2322	0.2627	0.2734
8	5	0.0000	0.0000	0.0004	0.0026	0.0092	0.0231	0.0467	0.0808	0.1239	0.1719	0.2188
8	6	0.0000	0.0000	0.0000	0.0002	0.0011	0.0038	0.0100	0.0217	0.0413	0.0703	0.1094
8	7	0.0000	0.0000	0.0000	0.0000	0.0001	0.0004	0.0012	0.0033	0.0079	0.0164	0.0313
8	8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0002	0.0007	0.0017	0.0039
9	0	0.9135	0.6302	0.3874	0.2316	0.1342	0.0751	0.0404	0.0207	0.0101	0.0046	0.0020
9	1	0.0830	0.2985	0.3874	0.3679	0.3020	0.2253	0.1556	0.1004	0.0605	0.0339	0.0176
9	2	0.0034	0.0629	0.1722	0.2597	0.3020	0.3003	0.2668	0.2162	0.1612	0.1110	0.0703
9	3	0.0001	0.0077	0.0446	0.1069	0.1762	0.2336	0.2668	0.2716	0.2508	0.2119	0.1641
9	4	0.0000	0.0006	0.0074	0.0283	0.0661	0.1168	0.1715	0.2194	0.2508	0.2600	0.2461
9	5	0.0000	0.0000	0.0008	0.0050	0.0165	0.0389	0.0735	0.1181	0.1672	0.2128	0.2461
9	6	0.0000	0.0000	0.0001	0.0006	0.0028	0.0087	0.0210	0.0424	0.0743	0.1160	0.1641
9	7	0.0000	0.0000	0.0000	0.0000	0.0003	0.0012	0.0039	0.0098	0.0212	0.0407	0.0703
9	8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0004	0.0013	0.0035	0.0083	0.0176
9	9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0003	0.0008	0.0020
10	0	0.9044	0.5987	0.3487	0.1969	0.1074	0.0563	0.0282	0.0135	0.0060	0.0025	0.0010
10	1	0.0914	0.3151	0.3874	0.3474	0.2684	0.1877	0.1211	0.0725	0.0403	0.0207	0.0098
10	2	0.0042	0.0746	0.1937	0.2759	0.3020	0.2816	0.2335	0.1757	0.1209	0.0763	0.0439
10	3	0.0001	0.0105	0.0574	0.1298	0.2013	0.2503	0.2668	0.2522	0.2150	0.1665	0.1172
10	4	0.0000	0.0010	0.0112	0.0401	0.0881	0.1460	0.2001	0.2377	0.2508	0.2384	0.2051
10	5	0.0000	0.0001	0.0015	0.0085	0.0264	0.0584	0.1029	0.1536	0.2007	0.2340	0.2461
10	6	0.0000	0.0000	0.0001	0.0012	0.0055	0.0162	0.0368	0.0689	0.1115	0.1596	0.2051
10	7	0.0000	0.0000	0.0000	0.0001	0.0008	0.0031	0.0090	0.0212	0.0425	0.0746	0.1172
10	8	0.0000	0.0000	0.0000	0.0000	0.0001	0.0004	0.0014	0.0043	0.0106	0.0229	0.0439
10	9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0005	0.0016	0.0042	0.0098
10	10	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0003	0.0010
11	0	0.8953	0.5688	0.3138	0.1673	0.0859	0.0422	0.0198	0.0088	0.0036	0.0014	0.0005
11	1	0.0995	0.3293	0.3835	0.3248	0.2362	0.1549	0.0932	0.0518	0.0266	0.0125	0.0054
11	2	0.0050	0.0867	0.2131	0.2866	0.2953	0.2581	0.1998	0.1395	0.0887	0.0513	0.0269
11	3	0.0002	0.0137	0.0710	0.1517	0.2215	0.2581	0.2568	0.2254	0.1774	0.1259	0.0806
11	4	0.0000	0.0014	0.0158	0.0536	0.1107	0.1721	0.2201	0.2428	0.2365	0.2060	0.1611
11	5	0.0000	0.0001	0.0025	0.0132	0.0388	0.0803	0.1321	0.1830	0.2207	0.2360	0.2256
11	6	0.0000	0.0000	0.0003	0.0023	0.0097	0.0268	0.0566	0.0985	0.1471	0.1931	0.2256
11	7	0.0000	0.0000	0.0000	0.0003	0.0017	0.0064	0.0173	0.0379	0.0701	0.1128	0.1611
11	8	0.0000	0.0000	0.0000	0.0000	0.0002	0.0011	0.0037	0.0102	0.0234	0.0462	0.0806

(*continúa*)

(*continuación*)

		Probabilidad										
		0.01	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
11	9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0005	0.0018	0.0052	0.0126	0.0269
11	10	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0007	0.0021	0.0054
11	11	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0005
12	0	0.8864	0.5404	0.2824	0.1422	0.0687	0.0317	0.0138	0.0057	0.0022	0.0008	0.0002
12	1	0.1074	0.3413	0.3766	0.3012	0.2062	0.1267	0.0712	0.0368	0.0174	0.0075	0.0029
12	2	0.0060	0.0988	0.2301	0.2924	0.2835	0.2323	0.1678	0.1088	0.0639	0.0339	0.0161
12	3	0.0002	0.0173	0.0852	0.1720	0.2362	0.2581	0.2397	0.1954	0.1419	0.0923	0.0537
12	4	0.0000	0.0021	0.0213	0.0683	0.1329	0.1936	0.2311	0.2367	0.2128	0.1700	0.1208
12	5	0.0000	0.0002	0.0038	0.0193	0.0532	0.1032	0.1585	0.2039	0.2270	0.2225	0.1934
12	6	0.0000	0.0000	0.0005	0.0040	0.0155	0.0401	0.0792	0.1281	0.1766	0.2124	0.2256
12	7	0.0000	0.0000	0.0000	0.0006	0.0033	0.0115	0.0291	0.0591	0.1009	0.1489	0.1934
12	8	0.0000	0.0000	0.0000	0.0001	0.0005	0.0024	0.0078	0.0199	0.0420	0.0762	0.1208
12	9	0.0000	0.0000	0.0000	0.0000	0.0001	0.0004	0.0015	0.0048	0.0125	0.0277	0.0537
12	10	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0008	0.0025	0.0068	0.0161
12	11	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0003	0.0010	0.0029
12	12	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0002
13	0	0.8775	0.5133	0.2542	0.1209	0.0550	0.0238	0.0097	0.0037	0.0013	0.0004	0.0001
13	1	0.1152	0.3512	0.3672	0.2774	0.1787	0.1029	0.0540	0.0259	0.0113	0.0045	0.0016
13	2	0.0070	0.1109	0.2448	0.2937	0.2680	0.2059	0.1388	0.0836	0.0453	0.0220	0.0095
13	3	0.0003	0.0214	0.0997	0.1900	0.2457	0.2517	0.2181	0.1651	0.1107	0.0660	0.0349
13	4	0.0000	0.0028	0.0277	0.0838	0.1535	0.2097	0.2337	0.2222	0.1845	0.1350	0.0873
13	5	0.0000	0.0003	0.0055	0.0266	0.0691	0.1258	0.1803	0.2154	0.2214	0.1989	0.1571
13	6	0.0000	0.0000	0.0008	0.0063	0.0230	0.0559	0.1030	0.1546	0.1968	0.2169	0.2095
13	7	0.0000	0.0000	0.0001	0.0011	0.0058	0.0186	0.0442	0.0833	0.1312	0.1775	0.2095
13	8	0.0000	0.0000	0.0000	0.0001	0.0011	0.0047	0.0142	0.0336	0.0656	0.1089	0.1571
13	9	0.0000	0.0000	0.0000	0.0000	0.0001	0.0009	0.0034	0.0101	0.0243	0.0495	0.0873
13	10	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0006	0.0022	0.0065	0.0162	0.0349
13	11	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0003	0.0012	0.0036	0.0095
13	12	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0005	0.0016
13	13	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001
14	0	0.8687	0.4877	0.2288	0.1028	0.0440	0.0178	0.0068	0.0024	0.0008	0.0002	0.0001
14	1	0.1229	0.3593	0.3559	0.2539	0.1539	0.0832	0.0407	0.0181	0.0073	0.0027	0.0009
14	2	0.0081	0.1229	0.2570	0.2912	0.2501	0.1802	0.1134	0.0634	0.0317	0.0141	0.0056
14	3	0.0003	0.0259	0.1142	0.2056	0.2501	0.2402	0.1943	0.1366	0.0845	0.0462	0.0222
14	4	0.0000	0.0037	0.0349	0.0998	0.1720	0.2202	0.2290	0.2022	0.1549	0.1040	0.0611
14	5	0.0000	0.0004	0.0078	0.0352	0.0860	0.1468	0.1963	0.2178	0.2066	0.1701	0.1222
14	6	0.0000	0.0000	0.0013	0.0093	0.0322	0.0734	0.1262	0.1759	0.2066	0.2088	0.1833
14	7	0.0000	0.0000	0.0002	0.0019	0.0092	0.0280	0.0618	0.1082	0.1574	0.1952	0.2095
14	8	0.0000	0.0000	0.0000	0.0003	0.0020	0.0082	0.0232	0.0510	0.0918	0.1398	0.1833
14	9	0.0000	0.0000	0.0000	0.0000	0.0003	0.0018	0.0066	0.0183	0.0408	0.0762	0.1222
14	10	0.0000	0.0000	0.0000	0.0000	0.0000	0.0003	0.0014	0.0049	0.0136	0.0312	0.0611
14	11	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0010	0.0033	0.0093	0.0222
14	12	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0005	0.0019	0.0056
14	13	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0002	0.0009
14	14	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001
15	0	0.8601	0.4633	0.2059	0.0874	0.0352	0.0134	0.0047	0.0016	0.0005	0.0001	0.0000

(*continúa*)

(*continuación*)

		Probabilidad										
		0.01	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
15	1	0.1303	0.3658	0.3432	0.2312	0.1319	0.0668	0.0305	0.0126	0.0047	0.0016	0.0005
15	2	0.0092	0.1348	0.2669	0.2856	0.2309	0.1559	0.0916	0.0476	0.0219	0.0090	0.0032
15	3	0.0004	0.0307	0.1285	0.2184	0.2501	0.2252	0.1700	0.1110	0.0634	0.0318	0.0139
15	4	0.0000	0.0049	0.0428	0.1156	0.1876	0.2252	0.2186	0.1792	0.1268	0.0780	0.0417
15	5	0.0000	0.0006	0.0105	0.0449	0.1032	0.1651	0.2061	0.2123	0.1859	0.1404	0.0916
15	6	0.0000	0.0000	0.0019	0.0132	0.0430	0.0917	0.1472	0.1906	0.2066	0.1914	0.1527
15	7	0.0000	0.0000	0.0003	0.0030	0.0138	0.0393	0.0811	0.1319	0.1771	0.2013	0.1964
15	8	0.0000	0.0000	0.0000	0.0005	0.0035	0.0131	0.0348	0.0710	0.1181	0.1647	0.1964
15	9	0.0000	0.0000	0.0000	0.0001	0.0007	0.0034	0.0116	0.0298	0.0612	0.1048	0.1527
15	10	0.0000	0.0000	0.0000	0.0000	0.0001	0.0007	0.0030	0.0096	0.0245	0.0515	0.0916
15	11	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0006	0.0024	0.0074	0.0191	0.0417
15	12	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0004	0.0016	0.0052	0.0139
15	13	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0003	0.0010	0.0032
15	14	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0005
15	15	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
16	0	0.8515	0.4401	0.1853	0.0743	0.0281	0.0100	0.0033	0.0010	0.0003	0.0001	0.0000
16	1	0.1376	0.3706	0.3294	0.2097	0.1126	0.0535	0.0228	0.0087	0.0030	0.0009	0.0002
16	2	0.0104	0.1463	0.2745	0.2775	0.2111	0.1336	0.0732	0.0353	0.0150	0.0056	0.0018
16	3	0.0005	0.0359	0.1423	0.2285	0.2463	0.2079	0.1465	0.0888	0.0468	0.0215	0.0085
16	4	0.0000	0.0061	0.0514	0.1311	0.2001	0.2252	0.2040	0.1553	0.1014	0.0572	0.0278
16	5	0.0000	0.0008	0.0137	0.0555	0.1201	0.1802	0.2099	0.2008	0.1623	0.1123	0.0667
16	6	0.0000	0.0001	0.0028	0.0180	0.0550	0.1101	0.1649	0.1982	0.1983	0.1684	0.1222
16	7	0.0000	0.0000	0.0004	0.0045	0.0197	0.0524	0.1010	0.1524	0.1889	0.1969	0.1746
16	8	0.0000	0.0000	0.0001	0.0009	0.0055	0.0197	0.0487	0.0923	0.1417	0.1812	0.1964
16	9	0.0000	0.0000	0.0000	0.0001	0.0012	0.0058	0.0185	0.0442	0.0840	0.1318	0.1746
16	10	0.0000	0.0000	0.0000	0.0000	0.0002	0.0014	0.0056	0.0167	0.0392	0.0755	0.1222
16	11	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0013	0.0049	0.0142	0.0337	0.0667
16	12	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0011	0.0040	0.0115	0.0278
16	13	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0008	0.0029	0.0085
16	14	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0005	0.0018
16	15	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0002
16	16	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
17	0	0.8429	0.4181	0.1668	0.0631	0.0225	0.0075	0.0023	0.0007	0.0002	0.0000	0.0000
17	1	0.1447	0.3741	0.3150	0.1893	0.0957	0.0426	0.0169	0.0060	0.0019	0.0005	0.0001
17	2	0.0117	0.1575	0.2800	0.2673	0.1914	0.1136	0.0581	0.0260	0.0102	0.0035	0.0010
17	3	0.0006	0.0415	0.1556	0.2359	0.2393	0.1893	0.1245	0.0701	0.0341	0.0144	0.0052
17	4	0.0000	0.0076	0.0605	0.1457	0.2093	0.2209	0.1868	0.1320	0.0796	0.0411	0.0182
17	5	0.0000	0.0010	0.0175	0.0668	0.1361	0.1914	0.2081	0.1849	0.1379	0.0875	0.0472
17	6	0.0000	0.0001	0.0039	0.0236	0.0680	0.1276	0.1784	0.1991	0.1839	0.1432	0.0944
17	7	0.0000	0.0000	0.0007	0.0065	0.0267	0.0668	0.1201	0.1685	0.1927	0.1841	0.1484
17	8	0.0000	0.0000	0.0001	0.0014	0.0084	0.0279	0.0644	0.1134	0.1606	0.1883	0.1855
17	9	0.0000	0.0000	0.0000	0.0003	0.0021	0.0093	0.0276	0.0611	0.1070	0.1540	0.1855
17	10	0.0000	0.0000	0.0000	0.0000	0.0004	0.0025	0.0095	0.0263	0.0571	0.1008	0.1484
17	11	0.0000	0.0000	0.0000	0.0000	0.0001	0.0005	0.0026	0.0090	0.0242	0.0525	0.0944
17	12	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0006	0.0024	0.0081	0.0215	0.0472
17	13	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0005	0.0021	0.0068	0.0182

(*continúa*)

(*continuación*)

		Probabilidad										
		0.01	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
17	14	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0004	0.0016	0.0052
17	15	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0003	0.0010
17	16	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001
17	17	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
18	0	0.8345	0.3972	0.1501	0.0536	0.0180	0.0056	0.0016	0.0004	0.0001	0.0000	0.0000
18	1	0.1517	0.3763	0.3002	0.1704	0.0811	0.0338	0.0126	0.0042	0.0012	0.0003	0.0001
18	2	0.0130	0.1683	0.2835	0.2556	0.1723	0.0958	0.0458	0.0190	0.0069	0.0022	0.0006
18	3	0.0007	0.0473	0.1680	0.2406	0.2297	0.1704	0.1046	0.0547	0.0246	0.0095	0.0031
18	4	0.0000	0.0093	0.0700	0.1592	0.2153	0.2130	0.1681	0.1104	0.0614	0.0291	0.0117
18	5	0.0000	0.0014	0.0218	0.0787	0.1507	0.1988	0.2017	0.1664	0.1146	0.0666	0.0327
18	6	0.0000	0.0002	0.0052	0.0301	0.0816	0.1436	0.1873	0.1941	0.1655	0.1181	0.0708
18	7	0.0000	0.0000	0.0010	0.0091	0.0350	0.0820	0.1376	0.1792	0.1892	0.1657	0.1214
18	8	0.0000	0.0000	0.0002	0.0022	0.0120	0.0376	0.0811	0.1327	0.1734	0.1864	0.1669
18	9	0.0000	0.0000	0.0000	0.0004	0.0033	0.0139	0.0386	0.0794	0.1284	0.1694	0.1855
18	10	0.0000	0.0000	0.0000	0.0001	0.0008	0.0042	0.0149	0.0385	0.0771	0.1248	0.1669
18	11	0.0000	0.0000	0.0000	0.0000	0.0001	0.0010	0.0046	0.0151	0.0374	0.0742	0.1214
18	12	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0012	0.0047	0.0145	0.0354	0.0708
18	13	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0012	0.0045	0.0134	0.0327
18	14	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0011	0.0039	0.0117
18	15	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0009	0.0031
18	16	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0006
18	17	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001
18	18	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
19	0	0.8262	0.3774	0.1351	0.0456	0.0144	0.0042	0.0011	0.0003	0.0001	0.0000	0.0000
19	1	0.1586	0.3774	0.2852	0.1529	0.0685	0.0268	0.0093	0.0029	0.0008	0.0002	0.0000
19	2	0.0144	0.1787	0.2852	0.2428	0.1540	0.0803	0.0358	0.0138	0.0046	0.0013	0.0003
19	3	0.0008	0.0533	0.1796	0.2428	0.2182	0.1517	0.0869	0.0422	0.0175	0.0062	0.0018
19	4	0.0000	0.0112	0.0798	0.1714	0.2182	0.2023	0.1491	0.0909	0.0467	0.0203	0.0074
19	5	0.0000	0.0018	0.0266	0.0907	0.1636	0.2023	0.1916	0.1468	0.0933	0.0497	0.0222
19	6	0.0000	0.0002	0.0069	0.0374	0.0955	0.1574	0.1916	0.1844	0.1451	0.0949	0.0518
19	7	0.0000	0.0000	0.0014	0.0122	0.0443	0.0974	0.1525	0.1844	0.1797	0.1443	0.0961
19	8	0.0000	0.0000	0.0002	0.0032	0.0166	0.0487	0.0981	0.1489	0.1797	0.1771	0.1442
19	9	0.0000	0.0000	0.0000	0.0007	0.0051	0.0198	0.0514	0.0980	0.1464	0.1771	0.1762
19	10	0.0000	0.0000	0.0000	0.0001	0.0013	0.0066	0.0220	0.0528	0.0976	0.1449	0.1762
19	11	0.0000	0.0000	0.0000	0.0000	0.0003	0.0018	0.0077	0.0233	0.0532	0.0970	0.1442
19	12	0.0000	0.0000	0.0000	0.0000	0.0000	0.0004	0.0022	0.0083	0.0237	0.0529	0.0961
19	13	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0005	0.0024	0.0085	0.0233	0.0518
19	14	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0006	0.0024	0.0082	0.0222
19	15	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0005	0.0022	0.0074
19	16	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0005	0.0018
19	17	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0003
19	18	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
19	19	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
20	0	0.8179	0.3585	0.1216	0.0388	0.0115	0.0032	0.0008	0.0002	0.0000	0.0000	0.0000
20	1	0.1652	0.3774	0.2702	0.1368	0.0576	0.0211	0.0068	0.0020	0.0005	0.0001	0.0000
20	2	0.0159	0.1887	0.2852	0.2293	0.1369	0.0669	0.0278	0.0100	0.0031	0.0008	0.0002

(*continúa*)

(*continuación*)

		Probabilidad										
		0.01	**0.05**	**0.10**	**0.15**	**0.20**	**0.25**	**0.30**	**0.35**	**0.40**	**0.45**	**0.50**
20	3	0.0010	0.0596	0.1901	0.2428	0.2054	0.1339	0.0716	0.0323	0.0123	0.0040	0.0011
20	4	0.0000	0.0133	0.0898	0.1821	0.2182	0.1897	0.1304	0.0738	0.0350	0.0139	0.0046
20	5	0.0000	0.0022	0.0319	0.1028	0.1746	0.2023	0.1789	0.1272	0.0746	0.0365	0.0148
20	6	0.0000	0.0003	0.0089	0.0454	0.1091	0.1686	0.1916	0.1712	0.1244	0.0746	0.0370
20	7	0.0000	0.0000	0.0020	0.0160	0.0545	0.1124	0.1643	0.1844	0.1659	0.1221	0.0739
20	8	0.0000	0.0000	0.0004	0.0046	0.0222	0.0609	0.1144	0.1614	0.1797	0.1623	0.1201
20	9	0.0000	0.0000	0.0001	0.0011	0.0074	0.0271	0.0654	0.1158	0.1597	0.1771	0.1602
20	10	0.0000	0.0000	0.0000	0.0002	0.0020	0.0099	0.0308	0.0686	0.1171	0.1593	0.1762
20	11	0.0000	0.0000	0.0000	0.0000	0.0005	0.0030	0.0120	0.0336	0.0710	0.1185	0.1602
20	12	0.0000	0.0000	0.0000	0.0000	0.0001	0.0008	0.0039	0.0136	0.0355	0.0727	0.1201
20	13	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0010	0.0045	0.0146	0.0366	0.0739
20	14	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0012	0.0049	0.0150	0.0370
20	15	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0003	0.0013	0.0049	0.0148
20	16	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0003	0.0013	0.0046
20	17	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0011
20	18	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002
20	19	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
20	20	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
21	0	0.8097	0.3406	0.1094	0.0329	0.0092	0.0024	0.0006	0.0001	0.0000	0.0000	0.0000
21	1	0.1718	0.3764	0.2553	0.1221	0.0484	0.0166	0.0050	0.0013	0.0003	0.0001	0.0000
21	2	0.0173	0.1981	0.2837	0.2155	0.1211	0.0555	0.0215	0.0072	0.0020	0.0005	0.0001
21	3	0.0011	0.0660	0.1996	0.2408	0.1917	0.1172	0.0585	0.0245	0.0086	0.0026	0.0006
21	4	0.0001	0.0156	0.0998	0.1912	0.2156	0.1757	0.1128	0.0593	0.0259	0.0095	0.0029
21	5	0.0000	0.0028	0.0377	0.1147	0.1833	0.1992	0.1643	0.1085	0.0588	0.0263	0.0097
21	6	0.0000	0.0004	0.0112	0.0540	0.1222	0.1770	0.1878	0.1558	0.1045	0.0574	0.0259
21	7	0.0000	0.0000	0.0027	0.0204	0.0655	0.1265	0.1725	0.1798	0.1493	0.1007	0.0554
21	8	0.0000	0.0000	0.0005	0.0063	0.0286	0.0738	0.1294	0.1694	0.1742	0.1442	0.0970
21	9	0.0000	0.0000	0.0001	0.0016	0.0103	0.0355	0.0801	0.1318	0.1677	0.1704	0.1402
21	10	0.0000	0.0000	0.0000	0.0003	0.0031	0.0142	0.0412	0.0851	0.1342	0.1673	0.1682
21	11	0.0000	0.0000	0.0000	0.0001	0.0008	0.0047	0.0176	0.0458	0.0895	0.1369	0.1682
21	12	0.0000	0.0000	0.0000	0.0000	0.0002	0.0013	0.0063	0.0206	0.0497	0.0933	0.1402
21	13	0.0000	0.0000	0.0000	0.0000	0.0000	0.0003	0.0019	0.0077	0.0229	0.0529	0.0970
21	14	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0005	0.0024	0.0087	0.0247	0.0554
21	15	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0006	0.0027	0.0094	0.0259
21	16	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0007	0.0029	0.0097
21	17	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0007	0.0029
21	18	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0006
21	19	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001
21	20	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
21	21	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
22	0	0.8016	0.3235	0.0985	0.0280	0.0074	0.0018	0.0004	0.0001	0.0000	0.0000	0.0000
22	1	0.1781	0.3746	0.2407	0.1087	0.0406	0.0131	0.0037	0.0009	0.0002	0.0000	0.0000
22	2	0.0189	0.2070	0.2808	0.2015	0.1065	0.0458	0.0166	0.0051	0.0014	0.0003	0.0001
22	3	0.0013	0.0726	0.2080	0.2370	0.1775	0.1017	0.0474	0.0184	0.0060	0.0016	0.0004
22	4	0.0001	0.0182	0.1098	0.1987	0.2108	0.1611	0.0965	0.0471	0.0190	0.0064	0.0017
22	5	0.0000	0.0034	0.0439	0.1262	0.1898	0.1933	0.1489	0.0913	0.0456	0.0187	0.0063

(*continúa*)

(*continuación*)

		Probabilidad										
		0.01	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
22	6	0.0000	0.0005	0.0138	0.0631	0.1344	0.1826	0.1808	0.1393	0.0862	0.0434	0.0178
22	7	0.0000	0.0001	0.0035	0.0255	0.0768	0.1391	0.1771	0.1714	0.1314	0.0812	0.0407
22	8	0.0000	0.0000	0.0007	0.0084	0.0360	0.0869	0.1423	0.1730	0.1642	0.1246	0.0762
22	9	0.0000	0.0000	0.0001	0.0023	0.0140	0.0451	0.0949	0.1449	0.1703	0.1586	0.1186
22	10	0.0000	0.0000	0.0000	0.0005	0.0046	0.0195	0.0529	0.1015	0.1476	0.1687	0.1542
22	11	0.0000	0.0000	0.0000	0.0001	0.0012	0.0071	0.0247	0.0596	0.1073	0.1506	0.1682
22	12	0.0000	0.0000	0.0000	0.0000	0.0003	0.0022	0.0097	0.0294	0.0656	0.1129	0.1542
22	13	0.0000	0.0000	0.0000	0.0000	0.0001	0.0006	0.0032	0.0122	0.0336	0.0711	0.1186
22	14	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0009	0.0042	0.0144	0.0374	0.0762
22	15	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0012	0.0051	0.0163	0.0407
22	16	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0003	0.0015	0.0058	0.0178
22	17	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0004	0.0017	0.0063
22	18	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0004	0.0017
22	19	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0004
22	20	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001
22	21	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
22	22	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
23	0	0.7936	0.3074	0.0886	0.0238	0.0059	0.0013	0.0003	0.0000	0.0000	0.0000	0.0000
23	1	0.1844	0.3721	0.2265	0.0966	0.0339	0.0103	0.0027	0.0006	0.0001	0.0000	0.0000
23	2	0.0205	0.2154	0.2768	0.1875	0.0933	0.0376	0.0127	0.0037	0.0009	0.0002	0.0000
23	3	0.0014	0.0794	0.2153	0.2317	0.1633	0.0878	0.0382	0.0138	0.0041	0.0010	0.0002
23	4	0.0001	0.0209	0.1196	0.2044	0.2042	0.1463	0.0818	0.0371	0.0138	0.0042	0.0011
23	5	0.0000	0.0042	0.0505	0.1371	0.1940	0.1853	0.1332	0.0758	0.0350	0.0132	0.0040
23	6	0.0000	0.0007	0.0168	0.0726	0.1455	0.1853	0.1712	0.1225	0.0700	0.0323	0.0120
23	7	0.0000	0.0001	0.0045	0.0311	0.0883	0.1500	0.1782	0.1602	0.1133	0.0642	0.0292
23	8	0.0000	0.0000	0.0010	0.0110	0.0442	0.1000	0.1527	0.1725	0.1511	0.1051	0.0584
23	9	0.0000	0.0000	0.0002	0.0032	0.0184	0.0555	0.1091	0.1548	0.1679	0.1433	0.0974
23	10	0.0000	0.0000	0.0000	0.0008	0.0064	0.0259	0.0655	0.1167	0.1567	0.1642	0.1364
23	11	0.0000	0.0000	0.0000	0.0002	0.0019	0.0102	0.0332	0.0743	0.1234	0.1587	0.1612
23	12	0.0000	0.0000	0.0000	0.0000	0.0005	0.0034	0.0142	0.0400	0.0823	0.1299	0.1612
23	13	0.0000	0.0000	0.0000	0.0000	0.0001	0.0010	0.0052	0.0182	0.0464	0.0899	0.1364
23	14	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0016	0.0070	0.0221	0.0525	0.0974
23	15	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0004	0.0023	0.0088	0.0258	0.0584
23	16	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0006	0.0029	0.0106	0.0292
23	17	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0008	0.0036	0.0120
23	18	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0010	0.0040
23	19	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0011
23	20	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002
23	21	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
23	22	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
23	23	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
24	0	0.7857	0.2920	0.0798	0.0202	0.0047	0.0010	0.0002	0.0000	0.0000	0.0000	0.0000
24	1	0.1905	0.3688	0.2127	0.0857	0.0283	0.0080	0.0020	0.0004	0.0001	0.0000	0.0000
24	2	0.0221	0.2232	0.2718	0.1739	0.0815	0.0308	0.0097	0.0026	0.0006	0.0001	0.0000
24	3	0.0016	0.0862	0.2215	0.2251	0.1493	0.0752	0.0305	0.0102	0.0028	0.0007	0.0001
24	4	0.0001	0.0238	0.1292	0.2085	0.1960	0.1316	0.0687	0.0289	0.0099	0.0028	0.0006

(*continúa*)

(*continuación*)

		Probabilidad										
		0.01	**0.05**	**0.10**	**0.15**	**0.20**	**0.25**	**0.30**	**0.35**	**0.40**	**0.45**	**0.50**
24	5	0.0000	0.0050	0.0574	0.1472	0.1960	0.1755	0.1177	0.0622	0.0265	0.0091	0.0025
24	6	0.0000	0.0008	0.0202	0.0822	0.1552	0.1853	0.1598	0.1061	0.0560	0.0237	0.0080
24	7	0.0000	0.0001	0.0058	0.0373	0.0998	0.1588	0.1761	0.1470	0.0960	0.0499	0.0206
24	8	0.0000	0.0000	0.0014	0.0140	0.0530	0.1125	0.1604	0.1682	0.1360	0.0867	0.0438
24	9	0.0000	0.0000	0.0003	0.0044	0.0236	0.0667	0.1222	0.1610	0.1612	0.1261	0.0779
24	10	0.0000	0.0000	0.0000	0.0012	0.0088	0.0333	0.0785	0.1300	0.1612	0.1548	0.1169
24	11	0.0000	0.0000	0.0000	0.0003	0.0028	0.0141	0.0428	0.0891	0.1367	0.1612	0.1488
24	12	0.0000	0.0000	0.0000	0.0000	0.0008	0.0051	0.0199	0.0520	0.0988	0.1429	0.1612
24	13	0.0000	0.0000	0.0000	0.0000	0.0002	0.0016	0.0079	0.0258	0.0608	0.1079	0.1488
24	14	0.0000	0.0000	0.0000	0.0000	0.0000	0.0004	0.0026	0.0109	0.0318	0.0694	0.1169
24	15	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0008	0.0039	0.0141	0.0378	0.0779
24	16	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0012	0.0053	0.0174	0.0438
24	17	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0003	0.0017	0.0067	0.0206
24	18	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0004	0.0021	0.0080
24	19	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0006	0.0025
24	20	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0006
24	21	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001
24	22	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
24	23	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
24	24	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
25	0	0.7778	0.2774	0.0718	0.0172	0.0038	0.0008	0.0001	0.0000	0.0000	0.0000	0.0000
25	1	0.1964	0.3650	0.1994	0.0759	0.0236	0.0063	0.0014	0.0003	0.0000	0.0000	0.0000
25	2	0.0238	0.2305	0.2659	0.1607	0.0708	0.0251	0.0074	0.0018	0.0004	0.0001	0.0000
25	3	0.0018	0.0930	0.2265	0.2174	0.1358	0.0641	0.0243	0.0076	0.0019	0.0004	0.0001
25	4	0.0001	0.0269	0.1384	0.2110	0.1867	0.1175	0.0572	0.0224	0.0071	0.0018	0.0004
25	5	0.0000	0.0060	0.0646	0.1564	0.1960	0.1645	0.1030	0.0506	0.0199	0.0063	0.0016
25	6	0.0000	0.0010	0.0239	0.0920	0.1633	0.1828	0.1472	0.0908	0.0442	0.0172	0.0053
25	7	0.0000	0.0001	0.0072	0.0441	0.1108	0.1654	0.1712	0.1327	0.0800	0.0381	0.0143
25	8	0.0000	0.0000	0.0018	0.0175	0.0623	0.1241	0.1651	0.1607	0.1200	0.0701	0.0322
25	9	0.0000	0.0000	0.0004	0.0058	0.0294	0.0781	0.1336	0.1635	0.1511	0.1084	0.0609
25	10	0.0000	0.0000	0.0001	0.0016	0.0118	0.0417	0.0916	0.1409	0.1612	0.1419	0.0974
25	11	0.0000	0.0000	0.0000	0.0004	0.0040	0.0189	0.0536	0.1034	0.1465	0.1583	0.1328
25	12	0.0000	0.0000	0.0000	0.0001	0.0012	0.0074	0.0268	0.0650	0.1140	0.1511	0.1550
25	13	0.0000	0.0000	0.0000	0.0000	0.0003	0.0025	0.0115	0.0350	0.0760	0.1236	0.1550
25	14	0.0000	0.0000	0.0000	0.0000	0.0001	0.0007	0.0042	0.0161	0.0434	0.0867	0.1328
25	15	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0013	0.0064	0.0212	0.0520	0.0974
25	16	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0004	0.0021	0.0088	0.0266	0.0609
25	17	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0006	0.0031	0.0115	0.0322
25	18	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0009	0.0042	0.0143
25	19	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0013	0.0053
25	20	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0003	0.0016
25	21	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0004
25	22	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001
25	23	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
25	24	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
25	25	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Apéndice 2

Tabla de probabilidades de Poisson

	x												
λ	0	1	2	3	4	5	6	7	8	9	10	11	12
0.1	0.9048	0.0905	0.0045	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.2	0.8187	0.1637	0.0164	0.0011	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.3	0.7408	0.2222	0.0333	0.0033	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.4	0.6703	0.2681	0.0536	0.0072	0.0007	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.5	0.6065	0.3033	0.0758	0.0126	0.0016	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.6	0.5488	0.3293	0.0988	0.0198	0.0030	0.0004	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.7	0.4966	0.3476	0.1217	0.0284	0.0050	0.0007	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.8	0.4493	0.3595	0.1438	0.0383	0.0077	0.0012	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.9	0.4066	0.3659	0.1647	0.0494	0.0111	0.0020	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1.0	0.3679	0.3679	0.1839	0.0613	0.0153	0.0031	0.0005	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
1.1	0.3329	0.3662	0.2014	0.0738	0.0203	0.0045	0.0008	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
1.2	0.3012	0.3614	0.2169	0.0867	0.0260	0.0062	0.0012	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000
1.3	0.2725	0.3543	0.2303	0.0998	0.0324	0.0084	0.0018	0.0003	0.0001	0.0000	0.0000	0.0000	0.0000
1.4	0.2466	0.3452	0.2417	0.1128	0.0395	0.0111	0.0026	0.0005	0.0001	0.0000	0.0000	0.0000	0.0000
1.5	0.2231	0.3347	0.2510	0.1255	0.0471	0.0141	0.0035	0.0008	0.0001	0.0000	0.0000	0.0000	0.0000
1.6	0.2019	0.3230	0.2584	0.1378	0.0551	0.0176	0.0047	0.0011	0.0002	0.0000	0.0000	0.0000	0.0000
1.7	0.1827	0.3106	0.2640	0.1496	0.0636	0.0216	0.0061	0.0015	0.0003	0.0001	0.0000	0.0000	0.0000
1.8	0.1653	0.2975	0.2678	0.1607	0.0723	0.0260	0.0078	0.0020	0.0005	0.0001	0.0000	0.0000	0.0000
1.9	0.1496	0.2842	0.2700	0.1710	0.0812	0.0309	0.0098	0.0027	0.0006	0.0001	0.0000	0.0000	0.0000
2.0	0.1353	0.2707	0.2707	0.1804	0.0902	0.0361	0.0120	0.0034	0.0009	0.0002	0.0000	0.0000	0.0000
2.1	0.1225	0.2572	0.2700	0.1890	0.0992	0.0417	0.0146	0.0044	0.0011	0.0003	0.0001	0.0000	0.0000
2.2	0.1108	0.2438	0.2681	0.1966	0.1082	0.0476	0.0174	0.0055	0.0015	0.0004	0.0001	0.0000	0.0000
2.3	0.1003	0.2306	0.2652	0.2033	0.1169	0.0538	0.0206	0.0068	0.0019	0.0005	0.0001	0.0000	0.0000
2.4	0.0907	0.2177	0.2613	0.2090	0.1254	0.0602	0.0241	0.0083	0.0025	0.0007	0.0002	0.0000	0.0000
2.5	0.0821	0.2052	0.2565	0.2138	0.1336	0.0668	0.0278	0.0099	0.0031	0.0009	0.0002	0.0000	0.0000
2.6	0.0743	0.1931	0.2510	0.2176	0.1414	0.0735	0.0319	0.0118	0.0038	0.0011	0.0003	0.0001	0.0000
2.7	0.0672	0.1815	0.2450	0.2205	0.1488	0.0804	0.0362	0.0139	0.0047	0.0014	0.0004	0.0001	0.0000
2.8	0.0608	0.1703	0.2384	0.2225	0.1557	0.0872	0.0407	0.0163	0.0057	0.0018	0.0005	0.0001	0.0000
2.9	0.0550	0.1596	0.2314	0.2237	0.1622	0.0940	0.0455	0.0188	0.0068	0.0022	0.0006	0.0002	0.0000
3.0	0.0498	0.1494	0.2240	0.2240	0.1680	0.1008	0.0504	0.0216	0.0081	0.0027	0.0008	0.0002	0.0001
3.1	0.0450	0.1397	0.2165	0.2237	0.1733	0.1075	0.0555	0.0246	0.0095	0.0033	0.0010	0.0003	0.0001
3.2	0.0408	0.1304	0.2087	0.2226	0.1781	0.1140	0.0608	0.0278	0.0111	0.0040	0.0013	0.0004	0.0001

(*continúa*)

(*continuación*)

	X												
λ	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**
0.1	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.3	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.4	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.5	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.6	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.7	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1.0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1.1	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1.2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1.3	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1.4	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1.5	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1.6	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1.7	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1.8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1.9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2.0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2.1	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2.2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2.3	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2.4	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2.5	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2.6	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2.7	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2.8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2.9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
3.0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
3.1	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
3.2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

(*continúa*)

(*continuación*)

	x												
λ	0	1	2	3	4	5	6	7	8	9	10	11	12
3.3	0.0369	0.1217	0.2008	0.2209	0.1823	0.1203	0.0662	0.0312	0.0129	0.0047	0.0016	0.0005	0.0001
3.4	0.0334	0.1135	0.1929	0.2186	0.1858	0.1264	0.0716	0.0348	0.0148	0.0056	0.0019	0.0006	0.0002
3.5	0.0302	0.1057	0.1850	0.2158	0.1888	0.1322	0.0771	0.0385	0.0169	0.0066	0.0023	0.0007	0.0002
3.6	0.0273	0.0984	0.1771	0.2125	0.1912	0.1377	0.0826	0.0425	0.0191	0.0076	0.0028	0.0009	0.0003
3.7	0.0247	0.0915	0.1692	0.2087	0.1931	0.1429	0.0881	0.0466	0.0215	0.0089	0.0033	0.0011	0.0003
3.8	0.0224	0.0850	0.1615	0.2046	0.1944	0.1477	0.0936	0.0508	0.0241	0.0102	0.0039	0.0013	0.0004
3.9	0.0202	0.0789	0.1539	0.2001	0.1951	0.1522	0.0989	0.0551	0.0269	0.0116	0.0045	0.0016	0.0005
4.0	0.0183	0.0733	0.1465	0.1954	0.1954	0.1563	0.1042	0.0595	0.0298	0.0132	0.0053	0.0019	0.0006
4.1	0.0166	0.0679	0.1393	0.1904	0.1951	0.1600	0.1093	0.0640	0.0328	0.0150	0.0061	0.0023	0.0008
4.2	0.0150	0.0630	0.1323	0.1852	0.1944	0.1633	0.1143	0.0686	0.0360	0.0168	0.0071	0.0027	0.0009
4.3	0.0136	0.0583	0.1254	0.1798	0.1933	0.1662	0.1191	0.0732	0.0393	0.0188	0.0081	0.0032	0.0011
4.4	0.0123	0.0540	0.1188	0.1743	0.1917	0.1687	0.1237	0.0778	0.0428	0.0209	0.0092	0.0037	0.0013
4.5	0.0111	0.0500	0.1125	0.1687	0.1898	0.1708	0.1281	0.0824	0.0463	0.0232	0.0104	0.0043	0.0016
4.6	0.0101	0.0462	0.1063	0.1631	0.1875	0.1725	0.1323	0.0869	0.0500	0.0255	0.0118	0.0049	0.0019
4.7	0.0091	0.0427	0.1005	0.1574	0.1849	0.1738	0.1362	0.0914	0.0537	0.0281	0.0132	0.0056	0.0022
4.8	0.0082	0.0395	0.0948	0.1517	0.1820	0.1747	0.1398	0.0959	0.0575	0.0307	0.0147	0.0064	0.0026
4.9	0.0074	0.0365	0.0894	0.1460	0.1789	0.1753	0.1432	0.1002	0.0614	0.0334	0.0164	0.0073	0.0030
5.0	0.0067	0.0337	0.0842	0.1404	0.1755	0.1755	0.1462	0.1044	0.0653	0.0363	0.0181	0.0082	0.0034
5.1	0.0061	0.0311	0.0793	0.1348	0.1719	0.1753	0.1490	0.1086	0.0692	0.0392	0.0200	0.0093	0.0039
5.2	0.0055	0.0287	0.0746	0.1293	0.1681	0.1748	0.1515	0.1125	0.0731	0.0423	0.0220	0.0104	0.0045
5.3	0.0050	0.0265	0.0701	0.1239	0.1641	0.1740	0.1537	0.1163	0.0771	0.0454	0.0241	0.0116	0.0051
5.4	0.0045	0.0244	0.0659	0.1185	0.1600	0.1728	0.1555	0.1200	0.0810	0.0486	0.0262	0.0129	0.0058
5.5	0.0041	0.0225	0.0618	0.1133	0.1558	0.1714	0.1571	0.1234	0.0849	0.0519	0.0285	0.0143	0.0065
5.6	0.0037	0.0207	0.0580	0.1082	0.1515	0.1697	0.1584	0.1267	0.0887	0.0552	0.0309	0.0157	0.0073
5.7	0.0033	0.0191	0.0544	0.1033	0.1472	0.1678	0.1594	0.1298	0.0925	0.0586	0.0334	0.0173	0.0082
5.8	0.0030	0.0176	0.0509	0.0985	0.1428	0.1656	0.1601	0.1326	0.0962	0.0620	0.0359	0.0190	0.0092
5.9	0.0027	0.0162	0.0477	0.0938	0.1383	0.1632	0.1605	0.1353	0.0998	0.0654	0.0386	0.0207	0.0102
6.0	0.0025	0.0149	0.0446	0.0892	0.1339	0.1606	0.1606	0.1377	0.1033	0.0688	0.0413	0.0225	0.0113
6.1	0.0022	0.0137	0.0417	0.0848	0.1294	0.1579	0.1605	0.1399	0.1066	0.0723	0.0441	0.0244	0.0124
6.2	0.0020	0.0126	0.0390	0.0806	0.1249	0.1549	0.1601	0.1418	0.1099	0.0757	0.0469	0.0265	0.0137
6.3	0.0018	0.0116	0.0364	0.0765	0.1205	0.1519	0.1595	0.1435	0.1130	0.0791	0.0498	0.0285	0.0150
6.4	0.0017	0.0106	0.0340	0.0726	0.1162	0.1487	0.1586	0.1450	0.1160	0.0825	0.0528	0.0307	0.0164
6.5	0.0015	0.0098	0.0318	0.0688	0.1118	0.1454	0.1575	0.1462	0.1188	0.0858	0.0558	0.0330	0.0179
6.6	0.0014	0.0090	0.0296	0.0652	0.1076	0.1420	0.1562	0.1472	0.1215	0.0891	0.0588	0.0353	0.0194
6.7	0.0012	0.0082	0.0276	0.0617	0.1034	0.1385	0.1546	0.1480	0.1240	0.0923	0.0618	0.0377	0.0210
6.8	0.0011	0.0076	0.0258	0.0584	0.0992	0.1349	0.1529	0.1486	0.1263	0.0954	0.0649	0.0401	0.0227

(*continúa*)

(*continuación*)

	X												
λ	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**
3.3	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
3.4	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
3.5	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
3.6	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
3.7	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
3.8	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
3.9	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
4.0	0.0002	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
4.1	0.0002	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
4.2	0.0003	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
4.3	0.0004	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
4.4	0.0005	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
4.5	0.0006	0.0002	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
4.6	0.0007	0.0002	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
4.7	0.0008	0.0003	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
4.8	0.0009	0.0003	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
4.9	0.0011	0.0004	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
5.0	0.0013	0.0005	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
5.1	0.0015	0.0006	0.0002	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
5.2	0.0018	0.0007	0.0002	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
5.3	0.0021	0.0008	0.0003	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
5.4	0.0024	0.0009	0.0003	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
5.5	0.0028	0.0011	0.0004	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
5.6	0.0032	0.0013	0.0005	0.0002	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
5.7	0.0036	0.0015	0.0006	0.0002	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
5.8	0.0041	0.0017	0.0007	0.0002	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
5.9	0.0046	0.0019	0.0008	0.0003	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
6.0	0.0052	0.0022	0.0009	0.0003	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
6.1	0.0058	0.0025	0.0010	0.0004	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
6.2	0.0065	0.0029	0.0012	0.0005	0.0002	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
6.3	0.0073	0.0033	0.0014	0.0005	0.0002	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
6.4	0.0081	0.0037	0.0016	0.0006	0.0002	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
6.5	0.0089	0.0041	0.0018	0.0007	0.0003	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
6.6	0.0099	0.0046	0.0020	0.0008	0.0003	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
6.7	0.0108	0.0052	0.0023	0.0010	0.0004	0.0001	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
6.8	0.0119	0.0058	0.0026	0.0011	0.0004	0.0002	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

(*continúa*)

(*continuación*)

	x												
λ	0	1	2	3	4	5	6	7	8	9	10	11	12
6.9	0.0010	0.0070	0.0240	0.0552	0.0952	0.1314	0.1511	0.1489	0.1284	0.0985	0.0679	0.0426	0.0245
7.0	0.0009	0.0064	0.0223	0.0521	0.0912	0.1277	0.1490	0.1490	0.1304	0.1014	0.0710	0.0452	0.0263
7.1	0.0008	0.0059	0.0208	0.0492	0.0874	0.1241	0.1468	0.1489	0.1321	0.1042	0.0740	0.0478	0.0283
7.2	0.0007	0.0054	0.0194	0.0464	0.0836	0.1204	0.1445	0.1486	0.1337	0.1070	0.0770	0.0504	0.0303
7.3	0.0007	0.0049	0.0180	0.0438	0.0799	0.1167	0.1420	0.1481	0.1351	0.1096	0.0800	0.0531	0.0323
7.4	0.0006	0.0045	0.0167	0.0413	0.0764	0.1130	0.1394	0.1474	0.1363	0.1121	0.0829	0.0558	0.0344
7.5	0.0006	0.0041	0.0156	0.0389	0.0729	0.1094	0.1367	0.1465	0.1373	0.1144	0.0858	0.0585	0.0366
7.6	0.0005	0.0038	0.0145	0.0366	0.0696	0.1057	0.1339	0.1454	0.1381	0.1167	0.0887	0.0613	0.0388
7.7	0.0005	0.0035	0.0134	0.0345	0.0663	0.1021	0.1311	0.1442	0.1388	0.1187	0.0914	0.0640	0.0411
7.8	0.0004	0.0032	0.0125	0.0324	0.0632	0.0986	0.1282	0.1428	0.1392	0.1207	0.0941	0.0667	0.0434
7.9	0.0004	0.0029	0.0116	0.0305	0.0602	0.0951	0.1252	0.1413	0.1395	0.1224	0.0967	0.0695	0.0457
8.0	0.0003	0.0027	0.0107	0.0286	0.0573	0.0916	0.1221	0.1396	0.1396	0.1241	0.0993	0.0722	0.0481
8.1	0.0003	0.0025	0.0100	0.0269	0.0544	0.0882	0.1191	0.1378	0.1395	0.1256	0.1017	0.0749	0.0505
8.2	0.0003	0.0023	0.0092	0.0252	0.0517	0.0849	0.1160	0.1358	0.1392	0.1269	0.1040	0.0776	0.0530
8.3	0.0002	0.0021	0.0086	0.0237	0.0491	0.0816	0.1128	0.1338	0.1388	0.1280	0.1063	0.0802	0.0555
8.4	0.0002	0.0019	0.0079	0.0222	0.0466	0.0784	0.1097	0.1317	0.1382	0.1290	0.1084	0.0828	0.0579
8.5	0.0002	0.0017	0.0074	0.0208	0.0443	0.0752	0.1066	0.1294	0.1375	0.1299	0.1104	0.0853	0.0604
8.6	0.0002	0.0016	0.0068	0.0195	0.0420	0.0722	0.1034	0.1271	0.1366	0.1306	0.1123	0.0878	0.0629
8.7	0.0002	0.0014	0.0063	0.0183	0.0398	0.0692	0.1003	0.1247	0.1356	0.1311	0.1140	0.0902	0.0654
8.8	0.0002	0.0013	0.0058	0.0171	0.0377	0.0663	0.0972	0.1222	0.1344	0.1315	0.1157	0.0925	0.0679
8.9	0.0001	0.0012	0.0054	0.0160	0.0357	0.0635	0.0941	0.1197	0.1332	0.1317	0.1172	0.0948	0.0703
9.0	0.0001	0.0011	0.0050	0.0150	0.0337	0.0607	0.0911	0.1171	0.1318	0.1318	0.1186	0.0970	0.0728
9.1	0.0001	0.0010	0.0046	0.0140	0.0319	0.0581	0.0881	0.1145	0.1302	0.1317	0.1198	0.0991	0.0752
9.2	0.0001	0.0009	0.0043	0.0131	0.0302	0.0555	0.0851	0.1118	0.1286	0.1315	0.1210	0.1012	0.0776
9.3	0.0001	0.0009	0.0040	0.0123	0.0285	0.0530	0.0822	0.1091	0.1269	0.1311	0.1219	0.1031	0.0799
9.4	0.0001	0.0008	0.0037	0.0115	0.0269	0.0506	0.0793	0.1064	0.1251	0.1306	0.1228	0.1049	0.0822
9.5	0.0001	0.0007	0.0034	0.0107	0.0254	0.0483	0.0764	0.1037	0.1232	0.1300	0.1235	0.1067	0.0844
9.6	0.0001	0.0007	0.0031	0.0100	0.0240	0.0460	0.0736	0.1010	0.1212	0.1293	0.1241	0.1083	0.0866
9.7	0.0001	0.0006	0.0029	0.0093	0.0226	0.0439	0.0709	0.0982	0.1191	0.1284	0.1245	0.1098	0.0888
9.8	0.0001	0.0005	0.0027	0.0087	0.0213	0.0418	0.0682	0.0955	0.1170	0.1274	0.1249	0.1112	0.0908
9.9	0.0001	0.0005	0.0025	0.0081	0.0201	0.0398	0.0656	0.0928	0.1148	0.1263	0.1250	0.1125	0.0928
10.0	0.0000	0.0005	0.0023	0.0076	0.0189	0.0378	0.0631	0.0901	0.1126	0.1251	0.1251	0.1137	0.0948

(*continúa*)

(*continuación*)

	x												
λ	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**
6.9	0.0130	0.0064	0.0029	0.0013	0.0005	0.0002	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
7.0	0.0142	0.0071	0.0033	0.0014	0.0006	0.0002	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
7.1	0.0154	0.0078	0.0037	0.0016	0.0007	0.0003	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
7.2	0.0168	0.0086	0.0041	0.0019	0.0008	0.0003	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
7.3	0.0181	0.0095	0.0046	0.0021	0.0009	0.0004	0.0001	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
7.4	0.0196	0.0104	0.0051	0.0024	0.0010	0.0004	0.0002	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
7.5	0.0211	0.0113	0.0057	0.0026	0.0012	0.0005	0.0002	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
7.6	0.0227	0.0123	0.0062	0.0030	0.0013	0.0006	0.0002	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
7.7	0.0243	0.0134	0.0069	0.0033	0.0015	0.0006	0.0003	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
7.8	0.0260	0.0145	0.0075	0.0037	0.0017	0.0007	0.0003	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
7.9	0.0278	0.0157	0.0083	0.0041	0.0019	0.0008	0.0003	0.0001	0.0001	0.0000	0.0000	0.0000	0.0000
8.0	0.0296	0.0169	0.0090	0.0045	0.0021	0.0009	0.0004	0.0002	0.0001	0.0000	0.0000	0.0000	0.0000
8.1	0.0315	0.0182	0.0098	0.0050	0.0024	0.0011	0.0005	0.0002	0.0001	0.0000	0.0000	0.0000	0.0000
8.2	0.0334	0.0196	0.0107	0.0055	0.0026	0.0012	0.0005	0.0002	0.0001	0.0000	0.0000	0.0000	0.0000
8.3	0.0354	0.0210	0.0116	0.0060	0.0029	0.0014	0.0006	0.0002	0.0001	0.0000	0.0000	0.0000	0.0000
8.4	0.0374	0.0225	0.0126	0.0066	0.0033	0.0015	0.0007	0.0003	0.0001	0.0000	0.0000	0.0000	0.0000
8.5	0.0395	0.0240	0.0136	0.0072	0.0036	0.0017	0.0008	0.0003	0.0001	0.0001	0.0000	0.0000	0.0000
8.6	0.0416	0.0256	0.0147	0.0079	0.0040	0.0019	0.0009	0.0004	0.0002	0.0001	0.0000	0.0000	0.0000
8.7	0.0438	0.0272	0.0158	0.0086	0.0044	0.0021	0.0010	0.0004	0.0002	0.0001	0.0000	0.0000	0.0000
8.8	0.0459	0.0289	0.0169	0.0093	0.0048	0.0024	0.0011	0.0005	0.0002	0.0001	0.0000	0.0000	0.0000
8.9	0.0481	0.0306	0.0182	0.0101	0.0053	0.0026	0.0012	0.0005	0.0002	0.0001	0.0000	0.0000	0.0000
9.0	0.0504	0.0324	0.0194	0.0109	0.0058	0.0029	0.0014	0.0006	0.0003	0.0001	0.0000	0.0000	0.0000
9.1	0.0526	0.0342	0.0208	0.0118	0.0063	0.0032	0.0015	0.0007	0.0003	0.0001	0.0000	0.0000	0.0000
9.2	0.0549	0.0361	0.0221	0.0127	0.0069	0.0035	0.0017	0.0008	0.0003	0.0001	0.0001	0.0000	0.0000
9.3	0.0572	0.0380	0.0235	0.0137	0.0075	0.0039	0.0019	0.0009	0.0004	0.0002	0.0001	0.0000	0.0000
9.4	0.0594	0.0399	0.0250	0.0147	0.0081	0.0042	0.0021	0.0010	0.0004	0.0002	0.0001	0.0000	0.0000
9.5	0.0617	0.0419	0.0265	0.0157	0.0088	0.0046	0.0023	0.0011	0.0005	0.0002	0.0001	0.0000	0.0000
9.6	0.0640	0.0439	0.0281	0.0168	0.0095	0.0051	0.0026	0.0012	0.0006	0.0002	0.0001	0.0000	0.0000
9.7	0.0662	0.0459	0.0297	0.0180	0.0103	0.0055	0.0028	0.0014	0.0006	0.0003	0.0001	0.0000	0.0000
9.8	0.0685	0.0479	0.0313	0.0192	0.0111	0.0060	0.0031	0.0015	0.0007	0.0003	0.0001	0.0001	0.0000
9.9	0.0707	0.0500	0.0330	0.0204	0.0119	0.0065	0.0034	0.0017	0.0008	0.0004	0.0002	0.0001	0.0000
10.0	0.0729	0.0521	0.0347	0.0217	0.0128	0.0071	0.0037	0.0019	0.0009	0.0004	0.0002	0.0001	0.0000

Apéndice 3

Tabla de áreas bajo la distribución *t* de Student

Distribución *t* de Student

	Probabilidad en el extremo				
Grados de libertad	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
40	1.303	1.684	2.021	2.423	2.704
50	1.299	1.676	2.009	2.403	2.678
60	1.296	1.671	2.000	2.390	2.660
70	1.294	1.667	1.994	2.381	2.648
80	1.292	1.664	1.990	2.374	2.639
90	1.291	1.662	1.987	2.368	2.632
100	1.290	1.660	1.984	2.364	2.626
110	1.289	1.659	1.982	2.361	2.621
120	1.289	1.658	1.980	2.358	2.617
130	1.288	1.657	1.978	2.355	2.614
140	1.288	1.656	1.977	2.353	2.611
150	1.287	1.655	1.976	2.351	2.609
160	1.287	1.654	1.975	2.350	2.607
170	1.287	1.654	1.974	2.348	2.605
180	1.286	1.653	1.973	2.347	2.603
190	1.286	1.653	1.973	2.346	2.602
200	1.286	1.653	1.972	2.345	2.601

Apéndice 4

Tablas de la distribución *F*

g.l.1g. 1.2	1	2	3	4	5	6	7	8	9	10	11
1	16 210.723	19 999.500	21 614.741	22 499.583	23 055.798	23 437.111	23 714.566	23 925.406	24 091.004	24 224.487	24 334.358
2	198.501	199.000	199.166	199.250	199.300	199.333	199.357	199.375	199.388	199.400	199.409
3	55.552	49.799	47.467	46.195	45.392	44.838	44.434	44.126	43.882	43.686	43.524
4	31.333	26.284	24.259	23.155	22.456	21.975	21.622	21.352	21.139	20.967	20.824
5	22.785	18.314	16.530	15.556	14.940	14.513	14.200	13.961	13.772	13.618	13.491
6	18.635	14.544	12.917	12.028	11.464	11.073	10.786	10.566	10.391	10.250	10.133
7	16.236	12.404	10.882	10.050	9.522	9.155	8.885	8.678	8.514	8.380	8.270
8	14.688	11.042	9.596	8.805	8.302	7.952	7.694	7.496	7.339	7.211	7.104
9	13.614	10.107	8.717	7.956	7.471	7.134	6.885	6.693	6.541	6.417	6.314
10	12.826	9.427	8.081	7.343	6.872	6.545	6.302	6.116	5.968	5.847	5.746
11	12.226	8.912	7.600	6.881	6.422	6.102	5.865	5.682	5.537	5.418	5.320
12	11.754	8.510	7.226	6.521	6.071	5.757	5.525	5.345	5.202	5.085	4.988
13	11.374	8.186	6.926	6.233	5.791	5.482	5.253	5.076	4.935	4.820	4.724
14	11.060	7.922	6.680	5.998	5.562	5.257	5.031	4.857	4.717	4.603	4.508
15	10.798	7.701	6.476	5.803	5.372	5.071	4.847	4.674	4.536	4.424	4.329
20	9.944	6.986	5.818	5.174	4.762	4.472	4.257	4.090	3.956	3.847	3.756
25	9.475	6.598	5.462	4.835	4.433	4.150	3.939	3.776	3.645	3.537	3.447
30	9.180	6.355	5.239	4.623	4.228	3.949	3.742	3.580	3.450	3.344	3.255
40	8.828	6.066	4.976	4.374	3.986	3.713	3.509	3.350	3.222	3.117	3.028
50	8.626	5.902	4.826	4.232	3.849	3.579	3.376	3.219	3.092	2.988	2.900
60	8.495	5.795	4.729	4.140	3.760	3.492	3.291	3.134	3.008	2.904	2.817
70	8.403	5.720	4.661	4.076	3.698	3.431	3.232	3.076	2.950	2.846	2.759
80	8.335	5.665	4.611	4.029	3.652	3.387	3.188	3.032	2.907	2.803	2.716
90	8.282	5.623	4.573	3.992	3.617	3.352	3.154	2.999	2.873	2.770	2.683
100	8.241	5.589	4.542	3.963	3.589	3.325	3.127	2.972	2.847	2.744	2.657
∞	7.879	5.298	4.279	3.715	3.350	3.091	2.897	2.744	2.621	2.519	2.432

g.l.1g. 1.2	12	13	14	15	20	25	30	40	50	60	70
1	24 426.366	24 504.536	24 571.767	24 630.205	24 835.971	24 960.340	25 043.628	25 148.153	25 211.089	25 253.137	25 283.216
2	199.416	199.423	199.428	199.433	199.450	199.460	199.466	199.475	199.480	199.483	199.485
3	43.387	43.271	43.172	43.085	42.778	42.591	42.466	42.308	42.213	42.149	42.104
4	20.705	20.603	20.515	20.438	20.167	20.002	19.892	19.752	19.667	19.611	19.570
5	13.384	13.293	13.215	13.146	12.903	12.755	12.656	12.530	12.454	12.402	12.366
6	10.034	9.950	9.877	9.814	9.589	9.451	9.358	9.241	9.170	9.122	9.088
7	8.176	8.097	8.028	7.968	7.754	7.623	7.534	7.422	7.354	7.309	7.276
8	7.015	6.938	6.872	6.814	6.608	6.482	6.396	6.288	6.222	6.177	6.145

(*continúa*)

(*continuación*)

g.l.1g. 1.2	12	13	14	15	20	25	30	40	50	60	70
9	6.227	6.153	6.089	6.032	5.832	5.708	5.625	5.519	5.454	5.410	5.379
10	5.661	5.589	5.526	5.471	5.274	5.153	5.071	4.966	4.902	4.859	4.828
11	5.236	5.165	5.103	5.049	4.855	4.736	4.654	4.551	4.488	4.445	4.414
12	4.906	4.836	4.775	4.721	4.530	4.412	4.331	4.228	4.165	4.123	4.092
13	4.643	4.573	4.513	4.460	4.270	4.153	4.073	3.970	3.908	3.866	3.835
14	4.428	4.359	4.299	4.247	4.059	3.942	3.862	3.760	3.698	3.655	3.625
15	4.250	4.181	4.122	4.070	3.883	3.766	3.687	3.585	3.523	3.480	3.450
20	3.678	3.611	3.553	3.502	3.318	3.203	3.123	3.022	2.959	2.916	2.885
25	3.370	3.304	3.247	3.196	3.013	2.898	2.819	2.716	2.652	2.609	2.577
30	3.179	3.113	3.056	3.006	2.823	2.708	2.628	2.524	2.459	2.415	2.383
40	2.953	2.888	2.831	2.781	2.598	2.482	2.401	2.296	2.230	2.184	2.150
50	2.825	2.760	2.703	2.653	2.470	2.353	2.272	2.164	2.097	2.050	2.015
60	2.742	2.677	2.620	2.570	2.387	2.270	2.187	2.079	2.010	1.962	1.927
70	2.684	2.619	2.563	2.513	2.329	2.211	2.128	2.019	1.949	1.900	1.864
80	2.641	2.577	2.520	2.470	2.286	2.168	2.084	1.974	1.903	1.854	1.817
90	2.608	2.544	2.487	2.437	2.253	2.134	2.051	1.939	1.868	1.818	1.781
100	2.583	2.518	2.461	2.411	2.227	2.108	2.024	1.912	1.840	1.790	1.752
∞	2.358	2.294	2.237	2.187	2.000	1.877	1.789	1.669	1.590	1.533	1.489

g.l.1g. l. 2	80	90	100	∞
1	25 305.799	25 323.378	25 337.450	25 464.075
2	199.487	199.488	199.490	199.498
3	42.070	42.043	42.022	41.828
4	19.540	19.516	19.497	19.325
5	12.338	12.317	12.300	12.144
6	9.062	9.042	9.026	8.879
7	7.251	7.232	7.217	7.076
8	6.121	6.103	6.088	5.951
9	5.356	5.337	5.322	5.188
10	4.805	4.787	4.772	4.639
11	4.391	4.373	4.359	4.226
12	4.069	4.051	4.037	3.904
13	3.812	3.794	3.780	3.647
14	3.602	3.584	3.569	3.436
15	3.427	3.409	3.394	3.260
20	2.861	2.843	2.828	2.690
25	2.553	2.534	2.519	2.377
30	2.358	2.339	2.323	2.176
40	2.125	2.105	2.088	1.932
50	1.989	1.968	1.951	1.786
60	1.900	1.878	1.861	1.689
70	1.837	1.815	1.797	1.618
80	1.789	1.767	1.748	1.563
90	1.752	1.730	1.711	1.520
100	1.723	1.700	1.681	1.485
∞	1.454	1.426	1.402	1.440

Apéndice 5

Tabla de áreas bajo la distribución ji cuadrada

	Probabilidad (porción de área en el extremo derecho)					
gl	**0.500**	**0.100**	**0.050**	**0.025**	**0.010**	**0.005**
1	0.455	2.706	3.841	5.024	6.635	7.879
2	1.386	4.605	5.991	7.378	9.210	10.597
3	2.366	6.251	7.815	9.348	11.345	12.838
4	3.357	7.779	9.488	11.143	13.277	14.860
5	4.351	9.236	11.070	12.833	15.086	16.750
6	5.348	10.645	12.592	14.449	16.812	18.548
7	6.346	12.017	14.067	16.013	18.475	20.278
8	7.344	13.362	15.507	17.535	20.090	21.955
9	8.343	14.684	16.919	19.023	21.666	23.589
10	9.342	15.987	18.307	20.483	23.209	25.188
11	10.341	17.275	19.675	21.920	24.725	26.757
12	11.340	18.549	21.026	23.337	26.217	28.300
13	12.340	19.812	22.362	24.736	27.688	29.819
14	13.339	21.064	23.685	26.119	29.141	31.319
15	14.339	22.307	24.996	27.488	30.578	32.801
16	15.338	23.542	26.296	28.845	32.000	34.267
17	16.338	24.769	27.587	30.191	33.409	35.718
18	17.338	25.989	28.869	31.526	34.805	37.156
19	18.338	27.204	30.144	32.852	36.191	38.582
20	19.337	28.412	31.410	34.170	37.566	39.997
21	20.337	29.615	32.671	35.479	38.932	41.401
22	21.337	30.813	33.924	36.781	40.289	42.796
23	22.337	32.007	35.172	38.076	41.638	44.181
24	23.337	33.196	36.415	39.364	42.980	45.559
25	24.337	34.382	37.652	40.646	44.314	46.928
26	25.336	35.563	38.885	41.923	45.642	48.290
27	26.336	36.741	40.113	43.195	46.963	49.645
28	27.336	37.916	41.337	44.461	48.278	50.993
29	28.336	39.087	42.557	45.722	49.588	52.336
30	29.336	40.256	43.773	46.979	50.892	53.672
31	30.336	41.422	44.985	48.232	52.191	55.003
32	31.336	42.585	46.194	49.480	53.486	56.328
33	32.336	43.745	47.400	50.725	54.776	57.648
34	33.336	44.903	48.602	51.966	56.061	58.964
35	34.336	46.059	49.802	53.203	57.342	60.275
36	35.336	47.212	50.998	54.437	58.619	61.581
37	36.336	48.363	52.192	55.668	59.893	62.883
38	37.335	49.513	53.384	56.896	61.162	64.181
39	38.335	50.660	54.572	58.120	62.428	65.476
40	39.335	51.805	55.758	59.342	63.691	66.766
50	49.335	63.167	67.505	71.420	76.154	79.490
60	59.335	74.397	79.082	83.298	88.379	91.952
70	69.334	85.527	90.531	95.023	100.425	104.215
80	79.334	96.578	101.879	106.629	112.329	116.321
90	89.334	107.565	113.145	118.136	124.116	128.299
100	99.334	118.498	124.342	129.561	135.807	140.169

Apéndice 6

Valores críticos de la *T* de Wilcoxon

De un extremo	De dos extremos	$n = 5$	$n = 6$	$n = 7$	$n = 8$	$n = 9$	$n = 10$
$\alpha = .05$	$\alpha = .10$	1	2	4	6	8	11
$\alpha = .025$	$\alpha = .05$		1	2	4	6	8
$\alpha = .01$	$\alpha = .02$			0	2	3	5
$\alpha = .005$	$\alpha = .01$				0	2	3
		$n = 11$	$n = 12$	$n = 13$	$n = 14$	$n = 15$	$n = 16$
$\alpha = .05$	$\alpha = .10$	14	17	21	26	30	36
$\alpha = .025$	$\alpha = .05$	11	14	17	21	25	30
$\alpha = .01$	$\alpha = .02$	7	10	13	16	20	24
$\alpha = .005$	$\alpha = .01$	5	7	10	13	16	19
		$n = 17$	$n = 18$	$n = 19$	$n = 20$	$n = 21$	$n = 22$
$\alpha = .05$	$\alpha = .10$	41	47	54	60	68	75
$\alpha = .025$	$\alpha = .05$	35	40	46	52	59	66
$\alpha = .01$	$\alpha = .02$	28	33	38	43	49	56
$\alpha = .005$	$\alpha = .01$	23	28	32	37	43	49
		$n = 23$	$n = 24$	$n = 25$	$n = 26$	$n = 27$	$n = 28$
$\alpha = .05$	$\alpha = .10$	83	92	101	110	120	130
$\alpha = .025$	$\alpha = .05$	73	81	90	98	107	117
$\alpha = .01$	$\alpha = .02$	62	69	77	85	93	102
$\alpha = .005$	$\alpha = .01$	55	61	68	76	84	92
		$n = 29$	$n = 30$	$n = 31$	$n = 32$	$n = 33$	$n = 34$
$\alpha = .05$	$\alpha = .10$	141	152	163	175	188	201
$\alpha = .025$	$\alpha = .05$	127	137	148	159	171	183
$\alpha = .01$	$\alpha = .02$	111	120	130	141	151	162
$\alpha = .005$	$\alpha = .01$	100	109	118	128	138	149
		$n = 35$	$n = 36$	$n = 37$	$n = 38$	$n = 39$	
$\alpha = .05$	$\alpha = .10$	214	228	242	256	271	
$\alpha = .025$	$\alpha = .05$	195	208	222	235	250	
$\alpha = .01$	$\alpha = .02$	174	186	198	211	224	
$\alpha = .005$	$\alpha = .01$	160	171	183	195	208	
		$n = 40$	$n = 41$	$n = 42$	$n = 43$	$n = 44$	$n = 45$
$\alpha = .05$	$\alpha = .10$	287	303	319	336	353	371
$\alpha = .025$	$\alpha = .05$	264	279	295	311	327	344
$\alpha = .01$	$\alpha = .02$	238	252	267	281	297	313
$\alpha = .005$	$\alpha = .01$	221	234	248	262	277	292
		$n = 46$	$n = 47$	$n = 48$	$n = 49$	$n = 50$	
$\alpha = .05$	$\alpha = .10$	389	408	427	446	466	
$\alpha = .025$	$\alpha = .05$	361	379	397	415	434	
$\alpha = .01$	$\alpha = .02$	329	345	362	380	398	
$\alpha = .005$	$\alpha = .01$	307	323	339	356	373	

Fuente: F. Wilcoxon y R.A. Wilcox, *Some Rapid Approximate Statistical Procedures*, 1964, p. 28. Reproducido con el permiso de American Cyanamid Company.

Fuente: Tomado de McClave, James T. y Benson George, *Statistics for Business and Economics*, 3a. edición, Dellen Publishing Company (San Francisco) y Collier Macmillan Publishers (London).

Apéndice 7

Tabla para la prueba de Mann Whitney de Daniel y Terrell

n_1	p	$n_2 = 2$	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2	0.001	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0.005	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
	0.01	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	2	2
	0.025	0	0	0	0	0	0	1	1	1	1	2	2	2	2	2	3	3	3	3
	0.05	0	0	0	0	1	1	2	2	2	2	3	3	4	4	4	4	5	5	5
	0.10	0	1	1	2	2	2	3	3	4	4	5	5	5	6	6	7	7	8	8
3	0.001	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1
	0.005	0	0	0	0	0	0	0	1	1	1	2	2	2	3	3	3	3	4	4
	0.01	0	0	0	0	0	1	1	2	2	2	3	3	3	4	4	5	5	5	6
	0.025	0	0	0	1	2	2	3	3	4	4	5	5	6	6	7	7	8	8	9
	0.05	0	1	1	2	3	3	4	5	5	6	6	7	8	8	9	10	10	11	12
	0.10	1	2	2	3	4	5	6	6	7	8	9	10	11	11	12	13	14	15	16
4	0.001	0	0	0	0	0	0	0	0	1	1	1	2	2	2	3	3	4	4	4
	0.005	0	0	0	0	1	1	2	2	3	3	4	4	5	6	6	7	7	7	9
	0.01	0	0	0	1	2	2	3	4	4	5	6	6	7	9	8	9	10	10	11
	0.025	0	0	1	2	3	4	5	5	6	7	8	9	10	11	12	12	13	14	15
	0.05	0	1	2	3	4	5	6	7	8	9	10	11	12	13	15	16	17	18	19
	0.10	1	2	4	5	6	7	8	10	11	12	13	14	16	17	18	19	21	22	23
5	0.001	0	0	0	0	0	0	1	2	2	3	3	4	4	5	6	6	7	8	8
	0.005	0	0	0	1	2	2	3	4	5	6	7	8	8	9	10	11	12	13	14
	0.01	0	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
	0.025	0	1	2	3	4	6	7	8	9	10	12	13	14	15	16	18	19	20	21
	0.05	1	2	3	5	6	7	9	10	12	13	14	16	17	19	20	21	23	24	26
	0.10	2	3	5	6	8	9	11	13	14	16	18	19	21	23	24	26	28	29	31
6	0.001	0	0	0	0	0	0	2	3	4	5	5	6	7	8	9	10	11	12	13
	0.005	0	0	1	2	3	4	5	6	7	8	10	11	12	13	14	16	17	18	19
	0.01	0	0	2	3	4	5	7	8	9	10	12	13	14	16	17	19	20	21	23
	0.025	0	2	3	4	6	7	9	11	12	14	15	17	18	20	22	23	25	26	28
	0.05	1	3	4	6	8	9	11	13	15	17	18	20	22	24	26	27	29	31	33
	0.10	2	4	6	8	10	12	14	16	18	20	22	24	26	28	30	32	35	37	39
7	0.001	0	0	0	0	1	2	3	4	6	7	8	9	10	11	12	14	15	16	17
	0.005	0	0	1	2	4	5	7	8	10	11	13	14	16	17	19	20	22	23	25
	0.01	0	1	2	4	5	7	8	10	12	13	15	17	18	20	22	24	25	27	29
	0.025	0	2	4	6	7	9	11	13	15	17	19	21	23	25	27	29	31	33	35
	0.05	1	3	5	7	9	12	14	16	18	20	22	25	27	29	31	34	36	38	40
	0.10	2	5	7	9	12	14	17	19	22	24	27	29	32	34	37	39	42	44	47
8	0.001	0	0	0	1	2	3	5	6	7	9	10	12	13	15	16	18	19	21	22
	0.005	0	0	2	3	5	7	8	10	12	14	16	18	19	21	23	25	27	29	31
	0.01	0	1	3	5	7	8	10	12	14	16	18	21	23	25	27	29	31	33	35
	0.025	1	3	5	7	9	11	14	16	18	20	23	25	27	30	32	35	37	39	42
	0.05	2	4	6	9	11	14	16	19	21	24	27	29	32	34	37	40	42	45	48
	0.10	3	6	8	11	14	17	20	23	25	28	31	34	37	40	43	46	49	52	55

(*continúa*)

(*continuación*)

n_1	p	$n_2 = 2$	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
9	0.001	0	0	0	2	3	4	6	8	9	11	13	15	16	18	20	22	24	26	27
	0.005	0	0	2	4	6	8	10	12	14	17	19	21	23	25	28	30	32	34	37
	0.01	0	1	3	6	8	10	12	15	17	19	22	24	27	29	32	34	37	39	41
	0.025	1	3	5	8	11	13	16	18	21	24	27	29	32	35	38	40	43	46	49
	0.05	2	5	7	10	13	16	19	22	25	28	31	34	37	40	43	46	49	52	55
	0.10	3	6	10	13	16	19	23	26	29	32	36	39	42	46	49	53	56	59	63
10	0.001	0	0	1	2	4	6	7	9	11	13	15	18	20	22	24	26	28	30	33
	0.005	0	1	3	5	7	10	12	14	17	19	22	25	27	30	32	35	38	40	43
	0.01	0	2	4	7	9	12	14	17	20	23	25	28	31	34	37	39	42	45	48
	0.025	1	4	6	9	12	15	18	21	24	27	30	34	37	40	43	46	49	53	56
	0.05	2	5	8	12	15	18	21	25	28	32	35	38	42	45	49	52	56	59	63
	0.10	4	7	11	14	18	22	25	29	33	37	40	44	48	52	55	59	63	67	71
11	0.001	0	0	1	3	5	7	9	11	13	16	18	21	23	25	28	30	33	35	38
	0.005	0	1	3	6	8	11	14	17	19	22	25	28	31	34	37	40	43	46	49
	0.01	0	2	5	8	10	13	16	19	23	26	29	32	35	38	42	45	48	51	54
	0.025	1	4	7	10	14	17	20	24	27	31	34	38	41	45	48	52	56	59	63
	0.05	2	6	9	13	17	20	24	28	32	35	39	43	47	51	55	58	62	66	70
	0.10	4	8	12	16	20	24	28	32	37	41	45	49	53	58	62	66	70	74	79
12	0.001	0	0	1	3	5	8	10	13	15	18	21	24	26	29	32	35	38	41	43
	0.005	0	2	4	7	10	13	16	19	22	25	28	32	35	38	42	45	48	52	55
	0.01	0	3	6	9	12	15	18	22	25	29	32	36	39	43	47	50	54	57	61
	0.025	2	5	8	12	15	19	23	27	30	34	38	42	46	50	54	58	62	66	70
	0.05	3	6	10	14	18	22	27	32	35	39	43	48	52	56	61	65	69	73	78
	0.1	5	9	13	18	22	27	31	36	40	45	50	54	59	64	68	73	78	82	87
13	0.001	0	0	2	4	6	9	12	15	18	21	24	27	30	33	36	39	43	46	49
	0.005	0	2	4	8	11	14	18	21	25	28	32	35	39	43	46	50	54	58	61
	0.01	1	3	6	10	13	17	21	24	28	32	36	40	44	48	52	56	60	64	68
	0.025	2	5	9	13	17	21	25	29	34	38	42	46	51	55	60	64	68	73	77
	0.05	3	7	11	16	20	25	29	34	38	43	48	52	57	62	66	71	76	81	85
	0.10	5	10	14	19	24	29	34	39	44	49	54	59	64	69	75	80	85	90	95
14	0.001	0	0	2	4	7	10	13	16	20	23	26	30	33	37	40	44	47	51	55
	0.005	0	2	5	8	12	16	19	23	27	31	35	39	43	47	51	55	59	64	68
	0.01	1	3	7	11	14	18	23	27	31	35	39	44	48	52	57	61	66	70	74
	0.025	2	6	10	14	18	23	27	32	37	41	46	51	56	60	65	70	75	79	84
	0.05	4	8	12	17	22	27	32	37	42	47	52	57	62	67	72	78	83	88	93
	0.10	5	11	16	21	26	32	37	42	48	53	59	64	70	75	81	86	92	98	103
15	0.001	0	0	2	5	8	11	15	18	22	25	29	33	37	41	44	41	52	56	60
	0.005	0	3	6	9	13	17	21	25	30	34	38	43	47	52	56	61	65	70	74
	0.01	1	4	8	12	16	20	25	29	34	38	43	48	52	57	62	67	71	76	81
	0.025	2	6	11	15	20	25	30	35	40	45	50	55	60	65	71	76	81	86	91
	0.05	4	8	13	19	24	29	34	40	45	51	56	62	67	73	78	84	89	95	101
	0.10	6	11	17	23	28	34	40	46	52	58	64	69	75	81	87	93	99	105	111
16	0.001	0	0	3	6	9	12	16	20	24	28	32	36	40	44	49	53	57	61	66
	0.005	0	3	6	10	14	19	23	28	32	37	42	46	51	56	61	66	71	75	80
	0.01	1	4	8	13	17	22	27	32	37	42	47	52	57	62	67	72	77	83	88
	0.025	2	7	12	16	22	27	32	38	43	48	54	60	65	71	76	82	87	93	99
	0.05	4	9	15	20	26	31	37	43	49	55	61	66	72	78	84	90	96	102	108
	0.10	6	12	18	24	30	37	43	49	55	62	68	75	81	87	94	100	107	113	120

(*continúa*)

(*continuación*)

n_1	p	$n_2 = 2$	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
17	0.001	0	1	3	6	10	14	18	22	26	30	35	39	44	48	53	58	62	67	71
	0.005	0	3	7	11	16	20	25	30	35	40	45	50	55	61	66	71	76	82	87
	0.01	1	5	9	14	19	24	29	34	39	45	50	56	61	67	72	78	83	89	94
	0.025	3	7	12	18	23	29	35	40	46	52	58	64	70	76	82	88	94	100	106
	0.05	4	10	16	21	27	34	40	46	52	58	65	71	78	84	90	97	103	110	116
	0.10	7	13	19	26	32	39	46	53	59	66	73	80	86	93	100	107	114	121	128
18	0.001	0	1	4	7	11	15	19	24	28	33	38	43	47	52	57	62	67	72	77
	0.005	0	3	7	12	17	22	27	32	38	43	48	54	59	65	71	76	82	88	93
	0.01	1	5	10	15	20	25	31	37	42	48	54	60	66	71	77	83	89	95	101
	0.025	3	8	13	19	25	31	37	43	49	56	62	68	75	81	87	94	100	107	113
	0.05	5	10	17	23	29	36	42	49	56	62	69	76	83	89	96	103	110	117	124
	0.10	7	14	21	28	35	42	49	56	63	70	78	85	92	99	107	114	121	129	136
19	0.001	0	1	4	8	12	16	21	26	30	35	41	46	51	56	61	67	72	78	83
	0.005	1	4	8	13	18	23	29	34	40	46	52	58	64	70	75	82	88	94	100
	0.01	2	5	10	16	21	27	33	39	45	51	57	64	70	76	83	89	95	102	108
	0.025	3	8	14	20	26	33	39	46	53	59	66	73	79	86	93	100	107	112	120
	0.05	5	11	18	24	31	38	45	52	59	66	73	81	88	95	102	110	117	124	131
	0.10	8	15	22	29	37	44	52	59	67	74	82	90	98	105	113	121	129	136	144
20	0.001	0	1	4	8	13	17	22	27	33	38	43	49	55	60	66	71	77	83	89
	0.005	1	4	9	14	19	25	31	37	43	49	55	61	68	74	80	87	93	100	106
	0.01	2	6	11	17	23	29	35	41	48	54	61	68	74	81	88	94	101	108	115
	0.025	3	9	15	21	28	35	42	49	56	63	70	77	84	91	99	106	113	120	128
	0.05	5	12	19	26	33	40	48	55	63	70	78	85	93	101	108	116	124	131	139
	0.10	8	16	23	31	39	47	55	63	71	79	87	95	103	111	120	128	136	144	152

Glosario

a

Análisis de regresión múltiple. Tiene solamente una variable dependiente y una independiente.
Análisis de regresión simple. Estudia la relación entre 2 variables.
Análisis de regresión y correlación. En economía, estudia la relación que pudiera existir entre la inversión y otras variables.
Análisis de series de tiempo. Comportamiento que tienen ciertos indicadores en el transcurso de un periodo.
Análisis de varianza (ANOVA). Es un conjunto de técnicas que se utilizan para probar hipótesis sobre la igualdad de más de 2 medias.
Análisis multivariado. Parte de la estadística que revisa varias técnicas empleadas en el análisis de más de 2 variables o mediciones.
Axioma. Es una proposición tan evidentemente cierta que no necesita demostración.

b

Bondad de ajuste. Grado de ajuste que tiene una serie de datos a una distribución normal, binomial o de otro tipo./ Prueba qué tan bien se ajustan los datos observados a determinada distribución teórica.

c

Coeficiente de determinación r 2. Es la razón de la variación explicada a la variación total.
Coeficiente de variación. Cociente entre la desviación estándar y la media aritmética multiplicado por 100.
Combinaciones. Son todos los subconjuntos de x elementos que se pueden formar de entre un conjunto de n objetos.
Complemento. El complemento de un conjunto A es el conjunto que se forma por todos los elementos que no pertenecen a ese conjunto.
Conglomerados. Subgrupos de una población que son relativamente pequeños (con pocos elementos) y numerosos.
Curtosis. Mide lo aplanado o puntiagudo de una distribución.
Curva característica operativa. Muestra los niveles de riesgo, las probabilidades de error al aceptar una hipótesis falsa para diversos valores hipotéticos de la verdadera media poblacional.

d

Datos. Materia prima de la estadística.
Datos continuos. Se pueden expresar con tal precisión que llega un momento en el que es difícil distinguir entre un número y el siguiente.
Datos de fuentes externas. Datos que se pueden obtener de otras personas u organizaciones.
Datos de fuentes internas. Datos que se generan al interior de la organización.
Datos de fuentes primarias. Datos que son generados por quien los utiliza, como sería el caso de los datos de fuentes internas.
Datos de fuentes secundarias. Datos que se obtienen de fuentes que no son los recopiladores originales de la información.
Datos discontinuos o discretos. Por su naturaleza se expresan en cantidades fácilmente distinguibles unas de otras.
Datos pareados. Datos dependientes.
Desviación estándar. Mide la dispersión de los datos alrededor de la media aritmética. / Raíz cuadrada de la varianza.
Desviación estándar de regresión. Es la desviación estándar de los valores y con respecto a los valores $\hat{y}$.
Desviación intercuartílica. Diferencia entre el tercer y el primer cuartiles.
Desviación media. Promedio de los valores absolutos de las diferencias entre cada dato y su media.
Distribución binomial. Distribución discreta de probabilidad que se basa en un experimento aleatorio.
Distribución de probabilidad para una variable aleatoria discreta. Tabla que enlista todos los valores que una variable puede asumir con su probabilidad de ocurrencia correspondiente.
Distribución leptocúrtica. Tiene un pico prominente al centro en comparación con la distribución normal.
Distribución mesocúrtica. Distribución tan puntiaguda o aplanada, según sea, como la distribución normal.
Distribución muestral. El conjunto de todas las muestras distintas de determinado tamaño n que es posible extraer de una población de tamaño N.
Distribución muestral de la media. Conjunto de las medias de todas las muestras de tamaño n que es posible obtener de una población de tamaño N.
Distribución normal o campana de Gauss. Distribución de probabilidad continua de amplia aplicación estadística y en otras disciplinas.

Distribución normal estándar. Tiene media aritmética de 0 y desviación estándar de 1.
Distribución platicúrtica. Tiene un pico relativamente bajo en comparación con la distribución normal.
Distribuciones de frecuencias. Series que utilizan frecuencias.

e

Ecuaciones normales. Conjunto de 2 ecuaciones que resueltas simultáneamente producen el valor de la pendiente y el de la ordenada al origen que son los parámetros de la ecuación lineal que arroja los mínimos cuadrados.
Error de muestreo. Diferencia entre el estadístico muestral que se utiliza para la inferencia y el valor verdadero del parámetro correspondiente.
Error tipo I. Error que consiste en rechazar una hipótesis que es verdadera.
Error tipo II. Error que consiste en aceptar una hipótesis falsa.
Errores que no se deben al muestreo. Son aquellos que se deben a la mala aplicación de los procedimientos.
Escala de intervalo. En ésta pueden realizarse todas las operaciones aritméticas (adición, sustracción, división y multiplicación).
Escala de razón. Tiene las capacidades de la escala de intervalo y además un cero absoluto que señala la carencia total de la característica que mide.
Escala nominal. La medición consiste en determinar si los casos específicos pertenecen a cierta categoría o no.
Escalas ordinales. En éstas pueden establecerse relaciones de mayor que o después de (>) y menor que o antes de (<).
Espacio muestral. Conjunto de todos los sucesos mutuamente excluyentes y colectivamente exhaustivos de un experimento aleatorio.
Estadística. Disciplina matemática considerada como un conjunto de técnicas para el análisis de datos.
Estadística descriptiva. Se ocupa del análisis de los datos sin utilizar muestras para hacer inferencias.
Estadística inferencial. Parte de la estadística que por medio del muestreo infiere conclusiones acerca de la totalidad de una población.
Estadístico muestral o estadístico. Medida de una muestra.
Estimación por intervalo. Utiliza un rango de valores o intervalo.
Estimación por punto. Utilizar un solo valor para estimar el parámetro.
Estimador. Cuando se utiliza un estadístico para estimar un parámetro, se dice que el estadístico se convierte en un estimador.
Estimador consistente. Conforme aumenta el tamaño de la muestra se incrementa la probabilidad de que el valor del estimador se aproxime al valor del parámetro.
Estimador eficiente. Es aquel con el menor error muestral, es decir, es aquel con la menor desviación estándar de su distribución muestral.
Estimador suficiente. Es aquel que agota toda la información relevante que se puede extraer de una muestra.
Estimar parámetros. Inferir el valor de la correspondiente medida de la población (parámetro) mediante una medida de una muestra (un estadístico).
Evento. Suceso o hecho de interés.
Eventos independientes. Suceden cuando la ocurrencia de uno de ellos no tiene efecto sobre la probabilidad de ocurrencia del otro.
Eventos mutuamente excluyentes. Eventos que no pueden ocurrir al mismo tiempo
Experimento aleatorio. Situaciones o ensayos que implican resultados inciertos.

f

Frecuencia relativa. Es la proporción de casos en cada categoría.

g

Gráficas circulares o de pastel. Gráficas que son útiles para visualizar la composición de un conjunto de datos.

h

Histogramas. Gráficas de barras en las que se usa un plano cartesiano (un plano con un eje vertical, el eje de las y y un eje horizontal, el eje de las x).
Hipótesis alternativa. Es la hipótesis que se asume como verdadera en caso de que la nula resulte no serlo.
Hipótesis estadística. Es una suposición o afirmación sobre alguna característica de una población.
Hipótesis nula. Es la hipótesis que desea probarse./ Es la afirmación o hipótesis que no es verdadera.
Homoscedasticidad. Propiedad que sostiene que para cada población de valor y_i le corresponderá una combinación de valor x_i.

i

Independencia estadística. Es la probabilidad condicional de la ocurrencia de un evento si la probabilidad de la de otro no tiene efectos sobre la suya.
Índice agregado. Es el índice que se calcula con los datos de varios valores, precios o cantidades.
Índice de Laspeyres. Se distingue porque utiliza las mismas cantidades del año base, para calcular los índices agregados de valor de ponderación fija.
Índice de Paasche. Se calcula utilizando, tanto en el numerador como en el denominador, las cantidades del año de interés.
Índice ideal de Fisher. Es la media geométrica de los índices de Laspeyres y de Paasche.

Índice simple. Es el índice calculado con los datos de un solo valor, precio o cantidad.
Ingreso. Marca la pendiente, lo cual lleva a la misma ecuación de regresión pero más precisa, con mayor número de decimales.
Intercepción. Es la misma ordenada al origen, es decir, el punto en el que la recta cruza (intercepta) al eje vertical.
Interpretación teórica o clásica de la probabilidad. Parte de que cuando no hay razones para preferir uno de los posibles resultados o sucesos, se considera que todos tienen la misma probabilidad de ocurrir.

l

Linealizar. Convertir la ecuación exponencial en una lineal para obtener la recta de mínimos cuadrados y volver la ecuación de regresión linealizada a su forma original (curva).

m

Marco muestral. La especificación de todos y cada uno de los elementos de la población a estudiar.
Media aritmética o promedio. Medida que se calcula sumando el total de los datos o valores de la variable para luego dividir esa suma entre el número de datos sumados.
Media armónica. Es el recíproco de la media aritmética de los recíprocos de los valores individuales.
Media geométrica. Raíz *n*-ésima del producto de los *n* datos o valores de la variable.
Media ponderada. Se utiliza principalmente para darle un peso relativo diferente a cada uno de los valores de la variable.
Mediana. Valor que ocupa el lugar central en una serie ordenada.
Medidas. Números simples que representan características de conjuntos de datos.
Método de mínimos cuadrados. Reduce al mínimo el cuadrado de las distancias verticales entre cada uno de los puntos y la recta ajustada.
Método del intervalo. Construcción de un intervalo que contenga el valor muestral observado de acuerdo con el valor de la media poblacional planteado en la hipótesis nula y el nivel de significación.
Método del valor de la P. Ayuda a comprender los procedimientos de prueba de hipótesis; reporta conclusiones sobre estudios estadísticos en publicaciones científicas.
Moda. Valor que más se repite, es decir, el que tiene mayor frecuencia.
Muestra. Subconjunto de los elementos de una población.
Muestra aleatoria simple. Es una sola muestra aleatoria.
Muestras aleatorias. Muestra representativa cuyos elementos son elegidos al azar. / Aquella en que todos los elementos de la población tienen o una probabilidad conocida de aparecer en una muestra o la misma para salir en ellas.
Muestras independientes. Muestras que se obtienen de poblaciones distintas.
Muestras múltiples. Es cuando se obtiene más de una muestra para el estudio.
Muestras relacionadas. Son mediciones diferentes de la misma muestra pero en condiciones diferentes.
Muestreo. Cantidad relativamente reducida de elementos representativos de una población.
Multicolinealidad. Es un problema que se da cuando las variables independientes están altamente correlacionadas.

n

Números índice. Valor relativo expresado en forma de porcentaje; mide precios, cantidades y valores durante un periodo dado contra el correspondiente precio en un periodo base. / Se utilizan para estudiar las variaciones que sufren determinadas mediciones de un periodo a otro, o durante diversos periodos.

p

Parámetro. Medida de una población.
Parámetros poblacionales. Medidas de las poblaciones.
Permutaciones. Son todos los subconjuntos de x elementos que se pueden formar de entre un conjunto de n objetos.
Población. Conjunto de todos los elementos o unidades de interés para un estudio determinado.
Precio productor. Precio fijado por el productor a la primera instancia compradora de su producto, excluyendo impuestos y costos de transportación facturados por separado.
Promedios móviles. Promedios que se calculan sucesivamente añadiendo cada vez el nuevo día y eliminando el más antiguo.
Proporción. Se calcula dividiendo el número de casos que tienen la característica de interés entre el total de elementos de la muestra (o de la población).
Pruebas de hipótesis. Procedimientos a través de los cuales se trata de verificar si ciertas suposiciones acerca de la población son ciertas o no. / Utilizar datos muestrales para evaluar la posible veracidad o precisión de suposiciones que se hacen sobre la población de interés.
Punto medio de clase o valor central de cada clase. Valor que representa a todos los de su clase.

r

Racha (o corrida). Es una sucesión ininterrumpida de casos del mismo tipo.
Rango. Diferencia entre el mayor y el menor de los valores; mide qué tan separados están los datos.

Región de rechazo. Son las 2 áreas que se encuentran en ambos extremos de la distribución.
Región de aceptación. Es el área que se encuentra alrededor de la media y limitada por los valores críticos de los extremos de la distribución.
Regla de las 5. Obliga a utilizar al menos 5 frecuencias esperadas en cada categoría.
Regla de Sturges. Primera aproximación al número de clases que debe tener la serie de clases y frecuencias mediante la raíz cuadrada del número de elementos.

S

Series. Conjuntos de datos que se presentan en tablas.
Series de datos agrupados. Tablas de datos en las que se resumen éstos de acuerdo con la frecuencia con la que se repiten o según determinados intervalos de valores.
Series de tiempo. Series de datos registrados en el tiempo que se trazan sobre el eje horizontal, en tanto que los valores de las observaciones se miden sobre el eje vertical. / Conjunto de observaciones de alguna variable tomadas a intervalos regulares.
Sesgo. Mide lo centrado o simétrico (sesgado) de una distribución.
Subconjunto. Un conjunto B es subconjunto de A, si todos los elementos de B son también elementos de A.

t

Tabla de datos no cruzados. Tabla cuyos datos de cada área son independientes entre sí.
Tasa constante. Cuando existe la misma probabilidad de que suceda un evento en cualquier momento dado.
Técnicas de análisis. Mecanismos mediante los cuales se convierten los datos en información útil.
Teorema de Chebyshev. Determina la proporción mínima de valores que se encuentran en un número específico de desviaciones estándar en relación con la media.
Teoría de la probabilidad. Se ocupa de analizar la forma en la que se miden diversos sucesos aleatorios.
Tratamiento. Es cualquier condición que se controla en el experimento.

U

Unidad de muestreo. Cada uno de los elementos a estudiar de una población.
Unidad experimental. Cada uno de los sujetos (persona, animal, corrida, sembrado o botella, por ejemplo) a los que se les aplica determinado tratamiento.

V

Valores corrientes. Se utilizan al momento sin descontar la inflación.
Valores reales. Son los valores corrientes ya con el ajuste de la inflación.
Variable. Característica que se mide al hacer determinadas observaciones y que puede asumir diferentes valores.
Variable aleatoria. Es aquella cuyo valor numérico se determina mediante el resultado de una situación incierta.
Variable binomial. Es aquella que sólo puede asumir 2 valores.
Variable continua. La diferencia entre un valor y el que le sigue es indistinguible. / Es aquella cuyos valores pueden medirse con tal precisión que la diferencia entre uno de sus valores y el siguiente puede perderse o ser insignificante.
Variable de respuesta. Es la medición que se realiza sobre las unidades experimentales.
Variable discontinua. Resulta clara la diferencia entre un valor y el que le sigue.
Variable discreta. Es aquella en la que pueden distinguirse sin lugar a dudas 2 valores contiguos.
Variable predictora. Es una variable independiente que se utiliza para hacer pronósticos sobre la variable dependiente.
Varianza. Mide la dispersión de los datos alrededor de la media aritmética. / Promedio de cuadrados.

Respuestas a los ejercicios nones

Capítulo 2 Tablas y gráficas

Ejercicios 2.2 Tablas

1.

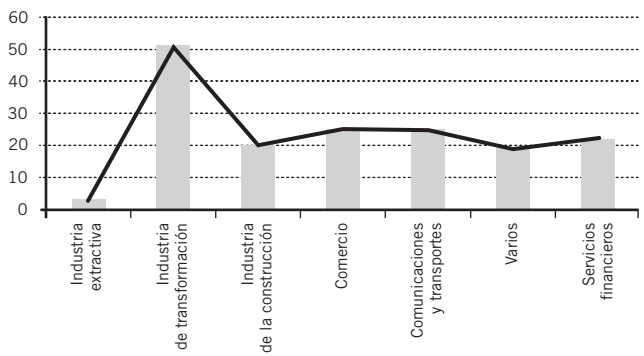
X Núm. de hijos	*f*	Frecuencias relativas	Frecuencias acumuladas
0	9	0.1286	9
1	7	0.1000	16
2	12	0.1714	28
3	10	0.1429	38
4	5	0.0714	43
5	6	0.0857	49
6	2	0.0286	51
7	5	0.0714	56
8	1	0.0143	57
9	5	0.0714	62
10	4	0.0571	66
11	1	0.0143	67
12	3	0.0429	70
	70	1.0000	

5.
a)

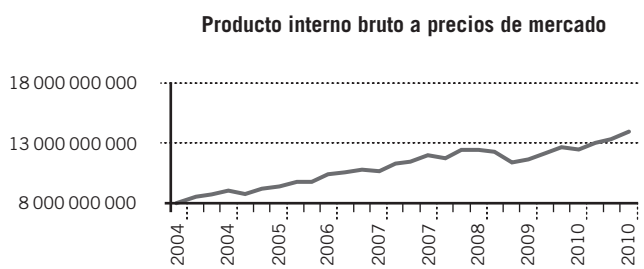
Tiempo de acceso (milisegundos)	Frecuencia *f*
4	1
7	1
8	1
11	2
12	20
13	1
14	4
15	15
16	28
17	16
18	2
19	3
20	3
21	2
22	3
23	1
24	15
25	2
29	1
$\sum f$	121

Ejercicios 2.3 Gráficas

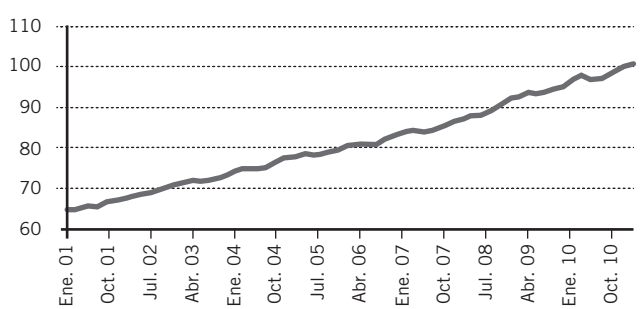
3.



5.

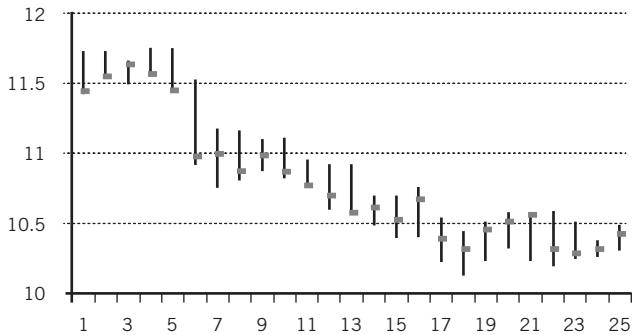


7.

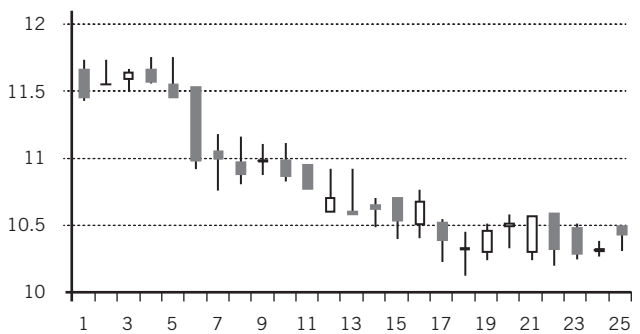


9.

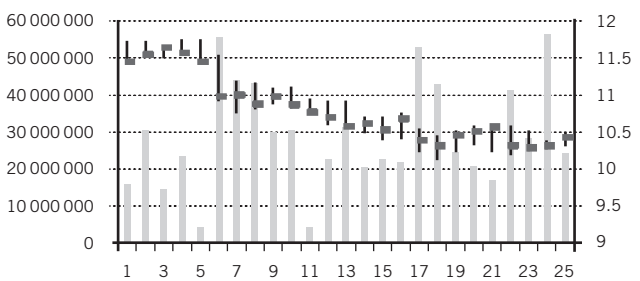
a) Máximo, mínimo y cierre



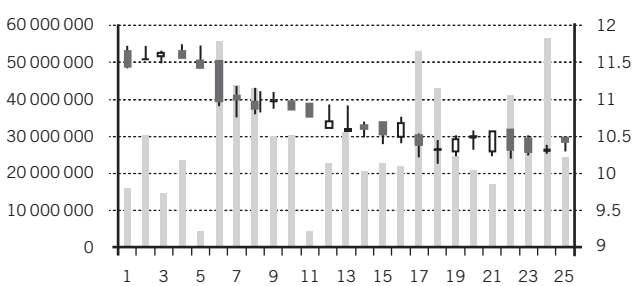
b) Apertura, máximo, mínimo y cierre, como velas japonesas



c) Volumen, máximo, mínimo y cierre



d) Volumen, apertura, máximo, mínimo y cierre, como velas japonesas



Capítulo 3 Medidas de posición, de dispersión, de composición y de forma

Ejercicios 3.1 Medidas de posición o de tendencia central

1. *a*) $\bar{X} = \dfrac{\sum X_i}{n} = \dfrac{111.44}{24} = 4.64$

   b) $Med = \dfrac{(4.49 + 4.5)}{2} = 4.495$

   c) $Moda = 4.49$

   d) $Q_1 = 4.4325$
   $Q_3 = 4.45$

3.

Temperaturas máximas

a) $\bar{X} = \dfrac{817}{27} = 30.26$

b) $Med = 31$

c) $Moda = 31$

d) $Q_1 = 27$
$Q_3 = 32$

Temperaturas mínimas

a) $\bar{X} = \dfrac{504}{27} = 18.67$

b) $Med = 18$

c) $Moda = 17, 23$

d) $Q_1 = 17$
$Q_3 = 23$

5. *a*) $\bar{X} = \dfrac{3\,463}{50} = 69.26$

   b) $Med = \dfrac{(69 + 74)}{2} = 71.5$

   c) $Moda = 43, 67, 85, 97$

   d) $Q_1 = 53.75$
   $Q_3 = 86.25$

7. *a*) $\bar{X} = 91.61$

   b) $Med = 91$

   c) $Moda = 94$

   d) $Q_1 = 85$
   $Q_3 = 99$

9. *a*) $\bar{X} = 3.626$

   b) $Med = 3$

   c) $Moda = 2$

   d) $Q_1 = 2$
   $Q_3 = 5$

11. *a*) $\bar{X} = 27.2$

    b) $Med = 26.571426$

    c) $Moda = 19.909$

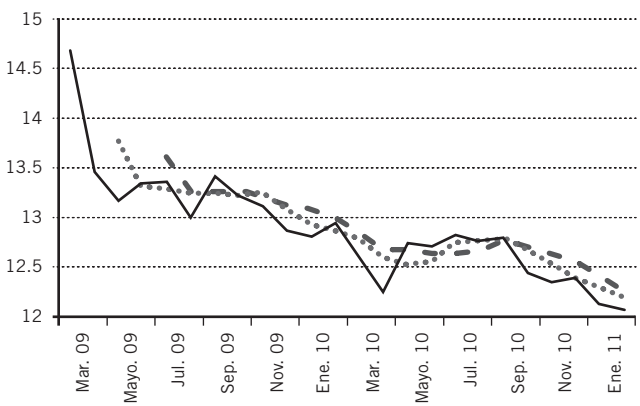
    d) $Q_1 = 20.16$
    $Q_3 = 65.50$

Nota: Estos datos pueden variar respecto a los que el estudiante obtenga debido a diferencias en la forma de construir la serie de clases y frecuencias.

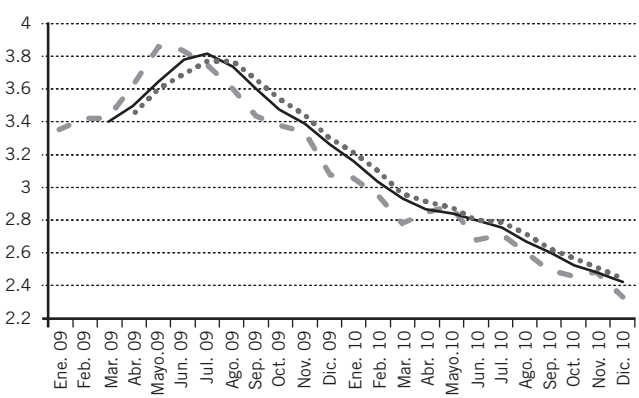
13. *a*) $\bar{X} = 143.76$
    b) $Med = 145.708$
    c) $Moda = 149.15$
    d) $Q_1 = 124.638$
    $Q_3 = 161.54$

15. *a*) $\bar{X} = 14.7$
    b) $Med = 14.8124$
    c) Esta serie tiene dos modas: 13 y 14
    d) $Q_1 = 11.9062$
    $Q_3 = 17.8214$

17. $\bar{X}_p = 8.51\%$
19. $\bar{X}_p = 135.32$
21. $\bar{X}_a = 83.33$
23. $\bar{X}_a = 69.44$
25. $\bar{X}_a = 4.5$
27. $\bar{X}_g = 1.8$
29. $\bar{X}_g = 28.04$
31.



33.



Ejercicios 3.2 Medidas de dispersión

1. *a*) Rango = 7
   b) DM = 2.08
   c) $s^2 = 5.84$
   d) $s^2 = 2.42$
   e) CV = 2.5

3. *a*) Rango = 4
   b) DM = 1.22
   c) $s^2 = 2.04$
   d) $s = 1.43$
   e) CV = 18.89

5. *a*) Rango = 447
   b) DM = 133.07
   c) $s^2 = 22\,520.98$
   d) $s = 150.07$
   e) CV = 28.80

7. *a*) Rango = 6
   b) DM = 1.18
   c) $s^2 = 2.3$
   d) $s = 1.51$
   e) CV = 65.09

9. *a*) Rango = 14
   b) DM = 3.198
   c) $s^2 = 14.40$
   d) $s = 3.79$
   e) CV = 46.18

11. *a*) Rango = 33
    b) DM = 6.61
    c) $s^2 = 64.28$
    d) $s = 8.02$
    e) CV = 41

13. *a*) Rango = 50
    b) DM = 10.88
    c) $s^2 = 164.23$
    d) $s = 12.81$
    e) CV = 43.94

15. *a*) Rango = 130
    b) DM = 23.65
    c) $s^2 = 852.39$
    d) $s = 29.19$
    e) CV = 41

17. 75% de los días las órdenes diarias están entre 224.27 y 321.83.
19. 75% de los días la sección de reptiles en el zoológico fue visitada entre 65.55 y 114.974 personas.

Ejercicios 3.4 Medidas de forma mediante momentos

1. *a*) $M_3 = -0.008418$
   $CS = -1.182$
   b) $M_4 = 0.004$
   $CK = -2.91$
3.

Temperaturas máximas

   a) $M_3 = 17.994631$
   $CS = 0.324$
   b) $M_4 = 713.1157$
   $CK = 0.3626$

Temperaturas mínimas

a) $M_3 = -34.982963$
$CS = -0.5076$
b) $M_4 = 647.3048025$
$CK = -0.7104$

5. *a*) $M_3 = -201.6753$
$CS = -0.2813$
b) $M_4 = 14\,511.045$
$CK = -0.7384$

7. *a*) $M_3 = 264.256$
$CS = 0.486$
b) $M_4 = 9\,184.8192$
$CK = -0.927$

Nota: Los resultados que se obtengan en este ejercicio pueden variar dependiendo de la forma en la que se construye la tabla de clases y frecuencias.

9. *a*) $M_3 = 163\,678.0667$

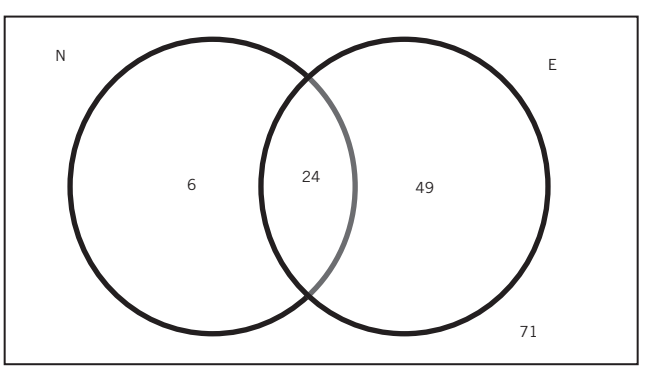
$$CS = \frac{M_3}{\sigma^3} = \frac{163\,678.0667}{\left(62.824^3\right)} = \frac{163\,678.0667}{247\,957.22} = 0.66$$

b) $M_4 = 9\,770\,938.54$
$CK = -2.373$

Capítulo 4 Introducción a la teoría de la probabilidad

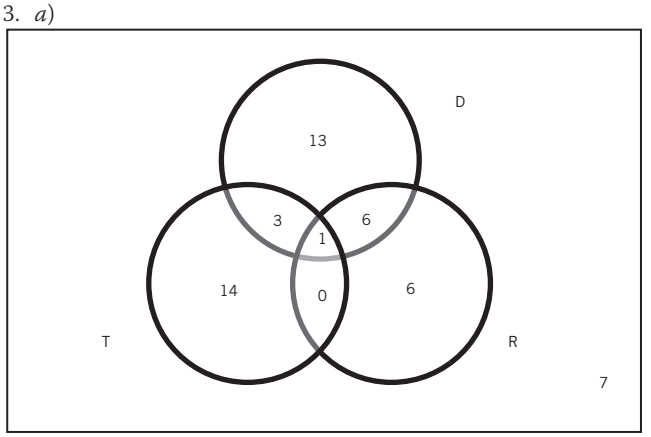
Ejercicios 4.1 Teoría de conjuntos y teoría de la probabilidad

1.

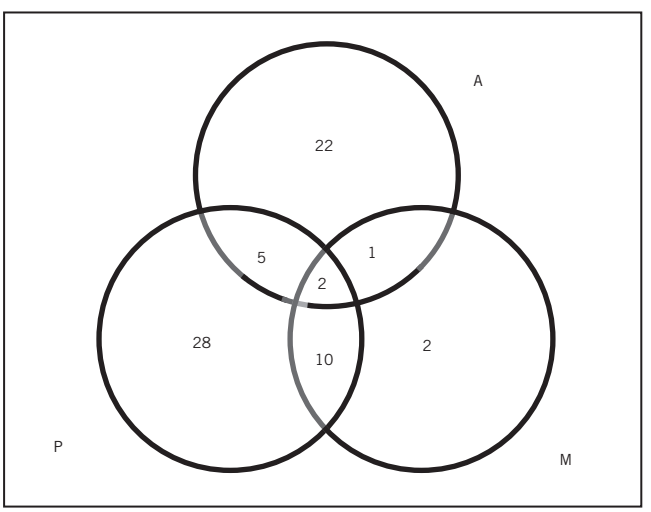


b) $n(E \cap N') = 49$
c) $n(N \cap E') = 6$

3. *a*)



5. *a*)



b) $n(A \cap (P \cup M)') = 22$
c) $n(P \cap (A \cup M)') = 25$
d) $n(M \cap (A \cup P)') = 2$
e) $n((A \cup P) \cap M') = 68$
f) $n((A \cap M) \cup P') = 42$
g) $n((M \cup P) \cap A') = 48$

7.
a) $n\,(M) = 140$
b) $n\,(no) = 85$
c) $n\,(si \cap H) = 115$
d) $n\,((si \cap M) \cup (no \cap H)) = 100 + 45 = 145$

Ejercicios 4.2 Conceptos básicos, terminología y notación

1.

Espacio muestral	Las 4 pelotas que se encuentran en el recipiente
Eventos posibles	Cualquiera de las 4 pelotas (azul, rojo, amarillo y naranja)

3.

Espacio muestral	Los 200 clientes de la tienda
Eventos posibles	Menor que 30, mayor que 30

5.

Espacio muestral	Los 160 hombres
Eventos posibles	Sí sabe, no sabe

7.

Espacio muestral	Las 400 personas encuestadas
Eventos posibles	(2, matutino), (2, vespertino), (2, nocturno), (4, matutino), (4, vespertino), (4, nocturno), (5, matutino), (5, vespertino), (5, nocturno), (7, matutino), (7, vespertino), (7, nocturno), (9, matutino), (9, vespertino), (9, nocturno), (11, matutino), (11, vespertino), (11, nocturno), (13, matutino), (12, vespertino), (13, nocturno)

Ejercicios 4.3 Técnicas de conteo, permutaciones y combinaciones

1. $k^n = 10^5 = 100\,000$
3. $10^4 = 10\,000$
5. Permutaciones. $P_n^x = \dfrac{n!}{(n-x)!} = \dfrac{5!}{(5-2)!} = \dfrac{5!}{3!} = 20$

   ae ai ao au ei eo eu io iu ou
   ea ia oa ua ei oe ue oi ui uo

   Combinaciones. $C_n^x = \dfrac{n!}{x!(n-x)!} = \dfrac{5!}{2!(5-2!)} = \dfrac{5!}{2!3!} = 10$

   ae ai ao au ei eo eu io iu ou

7. $P_n^x = \dfrac{n!}{(n-x)!} = \dfrac{10!}{(10-3)!} = \dfrac{10!}{7!} = 720$

9. $C_n^x = \dfrac{n!}{x!(n-x)!} = \dfrac{8!}{4!(8-4!)} = \dfrac{8!}{4!4!} = 70$

Ejercicios 4.4 Interpretaciones de la probabilidad

1. $P(Rota) = \dfrac{2}{25} = 0.08$

   $P(Completo) = \dfrac{23}{25} = 0.92$

3. *a)* $P(Mercadotecnia) = \dfrac{57}{130} = 0.44$

   b) $P(Finanzas) = \dfrac{24}{130} = 0.18$

5. *a)* $P(0) = \dfrac{1\,910}{2\,000} = 0.95$

   b) $P(Hasta\ 2) = \dfrac{1\,910 + 46 + 18}{2\,000} = \dfrac{1\,974}{2\,000} = 0.99$

7. $P(1) = \dfrac{30}{150} = \dfrac{1}{5}$

Ejercicios 4.5 Axiomas de la probabilidad

1. $P(\$1 \cup \$10) = P(\$1) + P(\$10) = \dfrac{25}{42} + \dfrac{1}{42} = \dfrac{26}{42} = \dfrac{13}{21}$

3. $P(pi \cup m) = P(pi) + P(m) = \dfrac{3}{45} + \dfrac{10}{45} = \dfrac{13}{45}$

5. *a)* $P(ciencia) = \dfrac{5}{25} = 0.2$

   b) $P(autos) = \dfrac{4}{25} = 0.16$

   c) $P(salud) = \dfrac{7}{25} = 0.28$

   d) $P(finanzas) = \dfrac{6}{25} = 0.24$

   e) $P(sociales) = \dfrac{3}{25} = 0.12$

   f) $P(c) + P(a) + P(s) + P(f) + P(so) = 0.2 + 0.16 + 0.28 + 0.24 + 0.12 = 1$ Sí cumplen

7.

Estado	Núm. de restaurantes	Frecuencia relativa
Distrito Federal	56	0.23
Hidalgo	53	0.22
Jalisco	43	0.18
Morelos	37	0.15
Puebla	28	0.11
Toluca	28	0.11
Total	245	1

Si cumplen $P(E_1) + P(E_2) + \cdots + P(E_n) = 1$

Ejercicios 4.6 Regla de la suma de probabilidades

1. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

   $= 0.33 + 0.27 - 0.1 = 0.5$

3. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

   $= 0.1 + 0.12 - 0.05 = 0.17$

5. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

   $= 0.7 + 0.8 - 0.6 = 0.9$

7. $p(n \cup i) = \dfrac{60}{110} + \dfrac{28}{110} - \dfrac{10}{110} - \dfrac{6}{110} - \dfrac{2}{110} + \dfrac{1}{110}$

   $= \dfrac{71}{110} = 0.71$

Ejercicios 4.7 Probabilidad condicional

1. *a*) $P(A|B) = \dfrac{P(A \text{ y } B)}{P(B)} = \dfrac{0.4}{0.5} = 0.80$

   b) $P(B|A) = \dfrac{P(A \text{ y } B)}{P(A)} = \dfrac{0.4}{0.6} = 0.67$

3. *a*) $P(C) = \dfrac{30}{100} = 0.3$

   $P(M \text{ y } C) = \dfrac{12}{100} = 0.12$

   $P(M|C) = \dfrac{P(M \text{ y } C)}{P(C)} = \dfrac{0.12}{0.3} = 0.4$

   b) $P(H) = \dfrac{47}{100} = 0.47$

   $P(Comodidad \text{ y } H) = \dfrac{14}{100} = 0.14$

   $P(Comodidad|H) = \dfrac{P(Comodidad \text{ y } H)}{P(H)}$

   $= \dfrac{0.14}{0.47} = 0.3$

5. $P(buen\ serv|precios\ altos)$

   $= \dfrac{P(precios\ altos\ y\ buen\ serv)}{P(precios\ altos)} = \dfrac{0.42}{0.6} = 0.7$

7. $P(tiempo|surten) = \dfrac{P(tiempo \text{ y } surten)}{P(tiempo)} = \dfrac{0.54}{0.72} = 0.75$

9. $P(l|N) = \dfrac{P(L \text{ y } N)}{P(N)} = \dfrac{0.35}{0.65} = 0.54$

Ejercicios 4.8 Independencia estadística

1. $(A \text{ y } B) = P(A) \cdot P(B) = (0.3)(0.1) = 0.03$

3. $P(A \text{ y } B) = P(A) \cdot P(B) = \dfrac{1}{2} \cdot \dfrac{1}{2} = \dfrac{1}{4}$

Son eventos independientes ya que el que caiga cara en la primera moneda no influye en lo que caiga en la segunda.

5. $P(A \text{ y } B) = P(A|B) \cdot P(B) = \dfrac{6}{30} \cdot \dfrac{5}{29} = \dfrac{30}{870} = \dfrac{1}{29}$

Son eventos dependientes.

Ejercicios 4.9 Regla de la multiplicación de probabilidades

1. $P(H \cap 100) = P(100) \cdot P(H|100)$

   $= \dfrac{18\ 475}{110\ 939\ 132} \cdot \dfrac{7\ 228}{18\ 475}$

   $= \dfrac{7\ 228}{110\ 939\ 132} = 0.000065$

3. *a*) $P(C \text{ y } P) = P(C)P(P|C) = (0.27)(0.19) = 0.053$

   b) $P(C \text{ y } L) = P(C)P(L|C) = (0.27)(0.22) = 0.059$

5. *a*) $P(P \text{ y } E) = P(P)P(E|P) = (0.84)(0.75) = 0.63$

   b) $P(P \text{ y } E) = P(P)P(NE|P) = (0.84)(0.25) = 0.21$

Ejercicios 4.10 Teorema de Bayes

1. $P(B|D) = 0.33$
3. $P(H_2|R) = 0.5$
5. $P(I_1|A) = 0.77$

Capítulo 5 Distribuciones de probabilidad discretas

Ejercicios 5.3 Media y varianza de una distribución de probabilidades

1. $E(X) = \mu = \sum X_i \cdot P(X_i) = 3.15$

   a) $Var(X) = \sigma^2 = \sum_{i=1}^{n} (X_i - \mu)^2 \cdot P(X_i) = 2.1275$

3. *a*) $E(X) = \mu = \sum X_i \cdot P(X_i) = 4.78$

   b) $Var(X) = \sigma^2 = \sum_{i=1}^{n} (X_i - \mu)^2 \cdot P(X_i) = 1.0716$

5. *a*) La probabilidad asociada a cada evento es:

Ventas (*X*)	Días (*f*)	P(*X*)
7	2	0.0667
8	4	0.1333
9	8	0.2667
10	7	0.2333
11	5	0.1667
12	4	0.1333
Sumas	30	1.0000

   b) $\mu = \sum X_i \cdot P(X_i) = 9.7$

   c) $\sigma^2 = \sum_{i=1}^{n} (X_i - \mu)^2 \cdot P(X_i) = 2.01$

Ejercicios 5.4 Distribución binomial

1. $P(X = 4|n = 6, p = 0.2) = 0.015$

3. *a*) $P(0) = 0.24$
   b) $P(2) = 0.2646$
   c) $P(2 \le X \le 4) = 0.3483$

5. *a*) $P(0) = 0.12$
   b) $P(0 \le X \le 3) = 0.26 + 0.28 + 0.19 = 0.73$
   c) $P(20) = 1 \times 10^{-20}$

La media y la varianza de una distribución de probabilidad binomial

7. *a*)

X_i	$P(X_i)$
0	0.59049
1	0.32805
2	0.0729
3	0.0081
4	0.00045
5	0.00001
	1

b) $\mu = 0.5$
c) $\sigma^2 = 0.45$

9. *a*)

X_i	$P(X_i)$
0	0.4304
1	0.4039
2	0.1421
3	0.0222
4	0.0013
	1

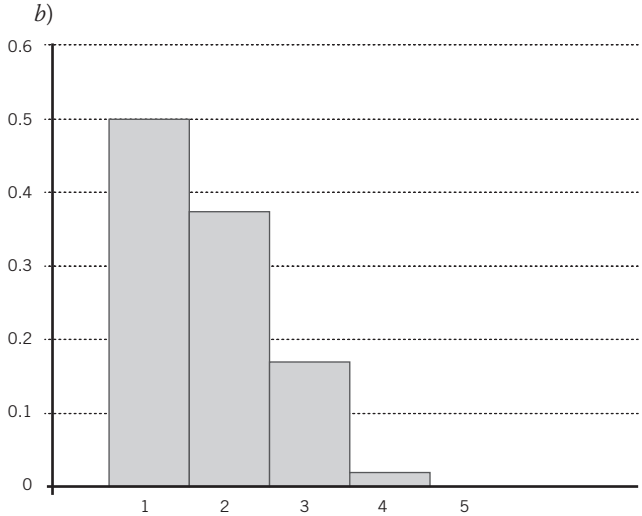
b) $\mu = 0.76$
c) $\sigma^2 = 0.615$

Ejercicios 5.5 Tres formas de presentar una distribución de probabilidad

1. *a*)

X_i	$P(X_i)$
0	0.4984
1	0.3723
2	0.1690
3	0.0166
4	0.0012
5	0.0002
	1

b)



c)

$$P(0|n = 5, p = 0.13) = \frac{5!}{0!(5-0)!}(0.13)^0(0.87)^5 = 0.4984$$

$$P(1|n = 5, p = 0.13) = \frac{5!}{1!(5-1)!}(0.13)^1(0.87)^4 = 0.3723$$

$$P(2|n = 5, p = 0.13) = \frac{5!}{2!(5-2)!}(0.13)^2(0.87)^3 = 0.1690$$

$$P(3|n = 5, p = 0.13) = \frac{5!}{3!(5-3)!}(0.13)^3(0.87)^2 = 0.0166$$

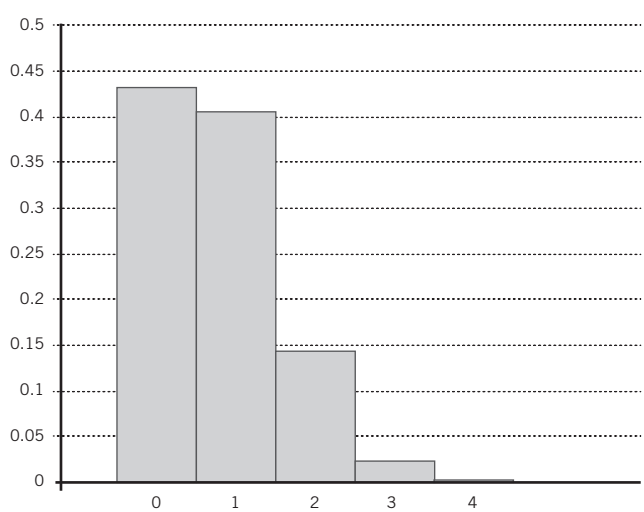
$$P(4|n = 5, p = 0.13) = \frac{5!}{4!(5-4)!}(0.13)^4(0.87)^1 = 0.0012$$

$$P(5|n = 5, p = 0.13) = \frac{5!}{5!(5-5)!}(0.13)^5(0.87)^0 = 0.0002$$

3. *a*)

X_i	$P(X_i)$
0	0.4304
1	0.4038
2	0.1421
3	0.0222
4	0.0013
	1.00

b)



c)

$$P(0|n = 4, p = 0.19) = \frac{4!}{0!(4-0)!}(0.19)^0(0.81)^4 = 0.4304$$

$$P(1|n = 4, p = 0.19) = \frac{4!}{1!(4-1)!}(0.19)^1(0.81)^3 = 0.4038$$

$$P(2|n = 4, p = 0.19) = \frac{4!}{2!(4-2)!}(0.19)^2(0.81)^2 = 0.1421$$

$$P(3|n = 4, p = 0.19) = \frac{4!}{3!(4-3)!}(0.19)^3(0.81)^1 = 0.0222$$

$$P(4|n = 4, p = 0.19) = \frac{4!}{4!(4-4)!}(0.19)^4(0.81)^0 = 0.0013$$

Ejercicios 5.6 Distribución de Poisson

1. $P(3) = 0.14$
3. $P(X > 2) = 1 - 0.81 - 0.16 - 0.02 = 0.01$
5. $P(0) = 0.67$

5.6.1 La distribución de Poisson como aproximación de la distribución binomial

7. *a*) y *b*) $P(4) = 0.0024$
9. *a*) y *b*) $P(7) = 0.0000163$

La media y la desviación estándar de la distribución de Poisson

11. $E(X) = Var(X) = np = 5$
13. $E(X) = Var(X) = np = 8$
15. $E(X) = Var(X) = np = 7$

Ejercicios 5.7 Distribución hipergeométrica

1. $P(2) = 0.42$
3. *a*) $P(0) = 0.17$
   b) $P(1) = 0.5$
   c) $P(2) = 0.3$
   d) $P(3) = 0.03$
5. *a*) $P(2) = 0.0008$
   b) $P(2) = 0.24$
   c) $P(5) = 0.005$
   d) $P(5) = 0.06$

La media y la desviación estándar de la distribución hipergeométrica

7. *a*)

Resultados (núm. de niños menores de tres años) X	$P(X)$
0	0.07142
1	0.3809
2	0.4285
3	0.1142
4	0.00476
$\sum P(X_i)$	1

b) y *c*) $\mu = 1.6$
$\sigma^2 = 0.64$

9. *a*)

Resultados (núm. de amigos capturados) X	$P(X)$
0	0.0004
1	0.0119
2	0.1008
3	0.3225
4	0.4032
5	0.1613
$\sum P(X_i)$	1.0000

b) $\mu = 3.60$
c) $\sigma = 0.84$

Ejercicios 5.8 Distribución multinomial

1. $P(X_1 = 2, X_2 = 1, X_3 = 2) = 0.14$
3. $P(X_1 = 2, X_2 = 3, X_3 = 5, X_4 = 1, X_5 = 4) = 0.3$
5. $P(X_1 = 1, X_2 = 1, X_3 = 1) = 0.04$

Media y desviación estándar de una distribución multinomial

7.

			Medias	Varianzas
	X_i	$p(X_i) = X_i/30$	$E(x_i) = np_i$	$Var(x_i) = np_i(1 - p_i)$
Fresa	13	0.4333	13.0000	7.3667
Limón	10	0.3333	10.0000	6.6667
Piña	7	0.2333	7.0000	5.3667
Total	30	1.0000	30.0000	

9.

			Medias	Varianzas
	X_i	$p(X_i) = X_i / 30$	$E(x_i) = np_i$	$Var(x_i) = np_i(1 - p_i)$
Gaseosas		0.43	86.0000	49.0200
Agua		0.2	40.0000	32.0000
Jugos		0.37	74.0000	46.6200
		1	200.0000	

Capítulo 6 Distribuciones de probabilidad continuas

Ejercicios 6.2.3 Tabla de áreas bajo la curva normal

1. $P(-1.75 \le z \le 2) = 0.9371$
3. $P(z \ge -1.25) = 0.8944$

5. $P(z \leq 2.74) = 0.9969$
7. $P(1.97 \leq z \leq 2.05) = 0.0042$
9. $P(z \leq -2.27) = 0.0116$
11. $z = 1.281$.
13. $z = 0.674$.
15. $z = 0.524$ y $z = 0.841$
17. $z = -0.722$ y $z = -0.628$
19. $z = 0.012$

Ejercicios 6.2.4 Determinación de probabilidades para cualquier distribución normal

1. *a*) 0.1151
   b) 0.8764
   c) 0.0107
3. *a*) 0.0051
   b) 0.0361
5. *a*) 0.0475
   b) 0.1596
7. $P(X \geq 3\,300\,000) = 10.56\%$
9. $P(25.5 \leq X \leq 26.5) = 24.27\%$
11. 15% de esa población ve 5.54 horas de televisión o más.
13. *a*) $\mu = 7.64$
    b) En 25% de los casos, el medicamento tarda, al menos, 10.336 horas en hacer efecto.
15. *a*) 106.7
    b) 99.8
17. 514.4

Ejercicios 6.3 Ajuste cuando se utiliza la distribución normal para evaluar probabilidades de una variable discreta (ajuste por discontinuidad)

1. $P = 0.3023$
3. $P = 0.0618$
5. $P = 0.0968$

Ejercicios 6.4 Aproximación de probabilidades de variables discontinuas con la distribución normal

6.4.1 Aproximación de la distribución binomial con la distribución normal

1. $P = 0.0202$
3. *a*) 0.0054
   b) 0.0129
   c) 0.9754
5. *a*) 0.9767
   b) 0.8621
7. *a*) 0.0216
   b) 0.0147
9. *a*) 0.2492
   b) 0.2422

Ejercicios 6.5 Distribución exponencial de probabilidad

1. 0.3935
3. *a*) 0.3935
   b) 0.7135
5. 0.1813

Capítulo 7 Muestreo y distribuciones muestrales

Ejercicios 7.2 Distribución muestral de la media

1. *a*) 17 310 309 460 000
   b) 499 500
   c) 252
   d) ∞
3. *a*) $\mu = \dfrac{\sum X}{N} = \dfrac{30}{5} = 2.5$

   b) $\sigma = \sqrt{\dfrac{\sum (X - \mu)^2}{N}} = \sqrt{\dfrac{17.5}{6}} = \sqrt{2.92} = 1.708$

   c) $C_n^N = \dfrac{N!}{n!(N-n)!} = \dfrac{6!}{3!(6-3)!} = 20$

Muestra	Defectos	$\bar{X}$	$(\bar{X}-\bar{X})$	$(\bar{X}-\bar{X})^2$
a, b, c	0, 1, 2	1	–1.5	2.25
a, b, d	0, 1, 3	1.33	–1.17	1.3689
a, b, e	0, 1, 4	1.67	–0.83	0.6889
a, b, f	0, 1, 5	2	–0.5	0.25
a, c, d	0, 2, 3	1.67	–0.83	0.6889
a, c, e	0, 2, 4	2	–0.5	0.25
a, c, f	0, 2, 5	2.33	–0.17	0.0289
a, d, e	0, 3, 4	2.33	–0.17	0.0289
a, d, f	0, 3, 5	2.67	0.17	0.0289
a, e, f	0, 4, 5	3	0.5	0.25
b, c, d	1, 2, 3	2	–0.5	0.25
b, c, e	1, 2, 4	2.33	–0.17	0.0289

(*continúa*)

(*continuación*)

Muestra	Defectos	$\bar{X}$	$(\bar{X}-\bar{X})$	$(\bar{X}-\bar{X})^2$
b, c, f	1, 2, 5	2.67	0.17	0.0289
b, d, e	1, 3, 4	2.67	0.17	0.0289
b, d, f	1, 3, 5	3	0.5	0.25
b, e, f	1, 4, 5	3.33	0.83	0.6889
c, d, e	2, 3, 4	3	0.5	0.25
c, d, f	2, 3, 5	3.33	0.83	0.6889
c, e, f	2, 4, 5	3.67	1.17	1.3689
d, e, f	3, 4, 5	4	1.5	2.25
Total		50	0	11.6668

d) $E(X) = \mu_{\bar{X}} = \dfrac{\sum \bar{X}}{n} = \dfrac{50}{20} = 2.5$

e) *1*

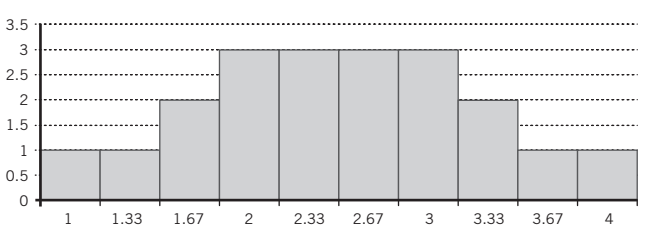
$$\sigma_{\bar{x}} = \sqrt{\frac{\sum(\bar{X} - \mu_{\bar{x}})^2}{n}} = \sqrt{\frac{11.6668}{20}} = \sqrt{05334} = 0.76$$

e) *2*

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \frac{1.708}{\sqrt{3}}\sqrt{\frac{6-3}{6-1}} = 0.76$$

f) 1 Sí
f) 2 Sí
g)

Media de la muestra	Frecuencia observada *f*
1	1
1.33	1
1.67	2
2	3
2.33	3
2.67	3
3	3
3.33	2
3.67	1
4	1



Se aproxima a la curva de la normal

h) $\mu \pm 3\sigma_{\bar{x}} = 2.5 \pm 3(0.76) = 2.5 \pm 2.28 = 0.22 - 4.78$

Todas las medias muestrales se encuentran dentro de este intervalo

5. $\sigma_{\bar{x}} = 7.19$

7. $\sigma_{\bar{x}} = 3.79$

9. *a*) $P(z \leq -2.78) = 0.5 - 0.4973 = 0.0027$
   b) $P(z \geq 1.11) = 0.5 - 0.3665 = 0.1335$
   c) $P(-1.67 \leq z \leq 1.67) = 0.4525 + 0.4525 = 0.9050$

11. *a*) $P(z \geq 2) = 0.5 - 0.4772 = 0.0228$
    b) $P(z \leq -1) = 0.5 - 0.3413 = 0.1587$
    c) $P(-0.67 \leq z \leq 0.67) = 0.2486 + 0.2486 = 0.4972$

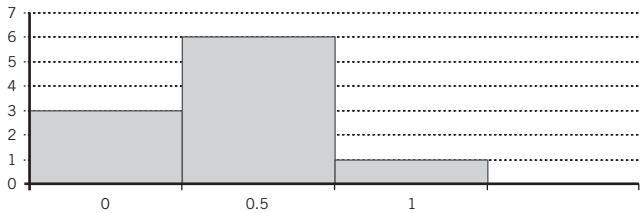
Ejercicios 7.3 Distribución muestral de la proporción

1. *a*) 0.4
   b) 0.4898
   c)

Muestras	*X*	Proporción (*p*)
a, b	1,1	$\frac{2}{2} = 1$
a, c	1,0	$\frac{1}{2} = 0.5$
a, d	1,0	$\frac{1}{2} = 0.5$
a, e	1,0	$\frac{1}{2} = 0.5$
b, c	1,0	$\frac{1}{2} = 0.5$
b, d	1,0	$\frac{1}{2} = 0.5$
b, e	1,0	$\frac{1}{2} = 0.5$
c, d	0,0	$\frac{0}{2} = 0$
c, e	0,0	$\frac{0}{2} = 0$
d, e	0,0	$\frac{0}{2} = 0$
Total		4

d) 0.4
e) 0.3
f) Sí

g)



h) Todas

3. 0.057387
5. 0.034641
7. Aumenta
9. *a*) 0.3707
   b) 0.0475
   c) 0.7387

Ejercicios 7.4 Distribución muestral de la varianza y estimadores sesgados e insesgados

1. *a*) 8
   b) 10
   c) 10
   d)

Muestra	Errores	$\bar{X}$	Varianzas modificadas
Alma, Susana	3, 5	4	2
Alma, Gabriela	3, 7	5	8
Alma, Corina	3, 9	6	18
Alma, Marisa	3, 11	7	32
Susana, Gabriela	5, 7	6	2
Susana, Corina	5, 9	7	8
Susana, Marisa	5, 11	8	18
Gabriela, Corina	7, 9	8	2
Gabriela, Marisa	7, 11	9	8
Corina, Marisa	9, 11	10	2
Total		70	100

   e) 5
   f) 10
   g) Sí
   h) Sí

3. Porque la media de la distribución muestral de la varianza no es igual que la varianza de la población.

Capítulo 8 Estimación de parámetros

Ejercicios 8.4 Estimación de una media para muestras grandes

5. \$131.01 a \$134.99
7. *a*) 42 906.4 a 43 093.6 kilómetros
   b) 41 726 a 44 274 kilómetros
9. 39.56 a 52.78 minutos
11. 9.96 a 28.14 horas
13. $n = 107$
15. $n = 17$
17. $n = 21$
19. $n = 118$
21. $n = 142$

Ejercicios 8.6 Estimación de una media con muestras pequeñas

1. 4.75 a 5.85
3. \$1 986.85 a \$2 247.15
5. 553.31 a 580.69
7. 28.97 a 35.03
9. \$499.34 a \$554.66

Ejercicios 8.7 Estimación de una proporción

1. 50 a 66%
3. 14 a 22%
5. 15.6 a 26.4%
7. $n = 107$
9. $n = 137$
11. $n = 121$
13. $n = 520$
15. $n = 127$

Ejercicios 8.8 Otros intervalos de confianza

1. \$4 997.09 a \$5 002.90
3. 3.2227 (3 años, 3 meses) a 3.7722 (3 años, 9 meses)
5. \$153.64 a \$166.36
7. 14.8% a 24.94%
9. 1.01% a 41.85%
11. \$65 505 a \$67 495
13. 8 413 450 000 a 8 786 550 000 kilómetros
15. 395 100 a 447 300 minutos. Es decir, 6 585 a 7 455 horas o de 274.38 a 310.63 días
17. 58 650 a 114 954
19. 1 300 128 a 1 495 147.2

Capítulo 9 Pruebas de hipótesis para una población

Ejercicios 9.6 Pruebas de uno y de dos extremos, y regiones de aceptación y de rechazo

1. Como el valor observado de la duración de las baterías, 805 horas, está fuera del intervalo de confianza $822.85 \geq \mu \leq 827.15$, se rechaza la hipótesis nula y se concluye que las baterías no duran, en promedio, 825 horas de uso continuo.
3. Como el valor observado de la duración del sabor de la goma de mascar, 15.7 minutos, está fuera del intervalo de

confianza $17.25 \geq \mu \leq 18.74$, se rechaza la hipótesis nula y se concluye que el sabor de la goma de mascar no dura, en promedio, 18 minutos.

5. Como el valor observado de la media, 5.2 cae dentro del intervalo de 3.87 a 5.7, se acepta H_0. El promedio que se tarda un cliente en realizar sus operaciones bancarias es de 4.8 minutos.
7. Como el valor observado de 113.9956 sale de la zona de aceptación de la hipótesis nula, se concluye, con la hipótesis alternativa, que el promedio de gasto por cliente es inferior a $121.
9. El límite para decir que las ventas se mantuvieron en el mismo promedio es de 59.9461 ventas diarias, pero como el promedio de la muestra rebasa esta cifra, rechazamos la hipótesis nula y se acepta que las ventas se incrementaron, es decir, se acepta la hipótesis alternativa.
11. Como el límite para considerar que el promedio es superior al normal es de 510 068.4821 y en este caso, la muestra nos dice que el promedio es de 510 000 entonces se acepta la hipótesis nula y se rechaza que este monto sea superior como lo señala la hipótesis alternativa.

Ejercicios 9.7 Tres métodos para realizar pruebas de hipótesis

1. **Método del estadístico de prueba**

   La z calculada es de -18.26, y está muy por debajo del límite que marca la z crítica, que es -1.96, por lo tanto no es posible aceptar la hipótesis nula.

   Método de P

   Se determina que la probabilidad de que z sea menor o igual que -18.26 es prácticamente de cero. Como esta probabilidad es mucho menor que el nivel de significación del 0.025 o 2.5% del extremo derecho, se rechaza la hipótesis nula.
3. **Método del estadístico de prueba**

   La z calculada es de -6.05 y la crítica es -1.96, como la primera se sale del límite que nos marca la segunda, entonces se rechaza la hipótesis nula y se concluye que el sabor de la goma de mascar no dura en promedio 18 minutos.

   Método de P

   Se determina que la probabilidad de que z sea menor o igual que -6.05 es prácticamente de cero. Como esta probabilidad es menor que el nivel de significación del 0.025, o 2.5% del extremo derecho, se rechaza la hipótesis nula y se concluye que el sabor de la goma de mascar no dura en promedio 18 minutos.
5. **Método del estadístico de prueba**

   La z calculada es de 1.11 y la crítica es 2.575, entonces se acepta la hipótesis nula porque la z calculada está dentro del límite de aceptación.

   Método de P

   Se determina que la probabilidad de que z sea mayor o igual que 1.11 es de $P(z \geq 1.11) = 0.1335$, o sea, 13.35%. Como esta probabilidad es mayor que el nivel de significación del 0.005, o 0.5% del extremo derecho, se acepta la hipótesis nula.
7. **Método del estadístico de prueba**

   La z calculada es de -1.559 y la crítica es -1.555, y aunque están muy cercanos, se rechaza la hipótesis nula porque la z calculada está fuera del límite de aceptación y se concluye, con la hipótesis alternativa, que el promedio de gasto por cliente es inferior a $121.

   Método de P

   Se determina que la probabilidad de que z sea mayor que 1.56 es de $P(z \geq 1.56) = 0.0594$, o sea, 5.94%. Como esta probabilidad es menor que el nivel de significación de 0.06 o 6% del extremo izquierdo, no se acepta la hipótesis nula.
9. **Método del estadístico de prueba**

   La z calculada es de 1.65 y la crítica es 1.645, entonces se rechaza la hipótesis nula porque esa z calculada está fuera del límite de aceptación y se admite que las ventas se incrementaron, es decir, se acepta la hipótesis alternativa.

   Método de P

   Se determina que la probabilidad de que z sea mayor que 1.65 es de $P(z \geq 1.65) = 0.0485$, o sea, 4.85%. Como esta probabilidad es menor que el nivel de significación del 0.05 o 5% del extremo derecho, se rechaza la hipótesis nula y se admite que las ventas se incrementaron, es decir, se acepta la hipótesis alternativa.
11. **Método del estadístico de prueba**

    La z calculada es de 2.04 y la crítica es 2.054, entonces no se rechaza la hipótesis nula porque la z calculada está dentro del límite de aceptación.

    Método de P

    Se determina que la probabilidad de que z sea mayor o igual que 2.04 es de $P(z \geq 2.04) = 0.0207$, o sea, 2.07%. Como esta probabilidad es mayor que el nivel de significación de 0.02, o 2% del extremo derecho, se acepta la hipótesis nula.

Ejercicios 9.8 Pruebas de hipótesis sobre una proporción poblacional

Pruebas de hipótesis para proporción

1. No es posible rechazar la hipótesis nula; se concluye que el porcentaje de las ventas del producto sigue siendo de 15% aun con la implementación de promociones por parte del gerente.
3. No se rechaza la hipótesis nula y se concluye que el procedimiento sugerido por el supervisor no disminuyó la cantidad de piezas defectuosas.
5. Se rechaza la hipótesis nula, llegando a la conclusión de que el porcentaje de cuentas con saldos mayores a $30 000 aumentó, lo cual quiere decir que las estrategias del banco funcionaron.

Capítulo 10 Pruebas de hipótesis para dos poblaciones

Ejercicios 10.2 Pruebas de hipótesis sobre la diferencia entre 2 medias

10.2.1 Pruebas con muestras grandes e independientes

1. El valor calculado de z cae dentro de la región de aceptación, por lo que no se rechaza la hipótesis nula y se concluye que no existe diferencia entre el nivel promedio de plomo en la sangre de los habitantes de la ciudad A y la ciudad B.
3. El valor calculado de z cae dentro de la región de rechazo, por lo que se rechaza la hipótesis nula y se concluye que existe una diferencia entre el promedio de veces que los alumnos de administración han consultado algún libro en la biblioteca y el promedio de los alumnos de contaduría.
5. Como el valor calculado de z cae en la región de aceptación, se acepta la hipótesis nula y se concluye que no existe una diferencia entre el tiempo promedio de minutos que hablan mensualmente los usuarios de cada tipo de servicio telefónico.
7. Como el valor calculado de z cae dentro de la región de rechazo, se rechaza la hipótesis nula y se concluye que el crecimiento promedio de los brotes de trigo a los que se les aplicó el fertilizante y el de los brotes a los que no, es diferente.
9. Como el valor calculado de z cae dentro de la región de rechazo, se rechaza la hipótesis nula y se concluye que el contenido promedio de nicotina de la marca nacional y de la marca extranjera es diferente.
11. Como el valor calculado de t está dentro del rango de aceptación, es decir, dentro de los valores críticos del estadístico de prueba, se acepta la hipótesis nula y se concluye que la resistencia del bloque de hormigón B es igual o mayor que la del bloque A.
13. Como el valor calculado de t es menor que el valor crítico determinado según el nivel de significación, –1.397, se rechaza la hipótesis nula y se concluye que es posible afirmar que el promedio de propinas recibidas en el turno vespertino es menor que en el matutino.
15. Como el valor calculado de t es menor que el valor crítico determinado según el nivel de significación, –1.684, se rechaza la hipótesis nula y se concluye que, efectivamente, el ingreso promedio de los miembros de la ciudad B es menor que el de los miembros de la ciudad A.
17. Como el valor calculado de t es mayor que el valor crítico, se rechaza la hipótesis nula y, de acuerdo con la hipótesis alternativa, se concluye que la duración de la batería A es mayor que la de la batería B.
19. Como el valor calculado de t es menor que el valor crítico, se rechaza la hipótesis nula y se concluye que el promedio de horas que las niñas ven televisión es menor que el de los niños.
21. Como el valor calculado de t es mayor que su valor crítico, se rechaza la hipótesis nula y se concluye que sí existe diferencia entre las calificaciones que los estudiantes respondieron en la encuesta y las que se tienen registradas en los archivos escolares.
23. Como el valor calculado de t cae en la región de rechazo, ya que es menor que el valor crítico mínimo, se rechaza la hipótesis nula y se concluye que sí existe diferencia entre las mediciones de los termómetros de tierra y los aéreos.

Ejercicios 10.3 Pruebas de hipótesis sobre la diferencia entre 2 proporciones

1. Como este valor calculado está muy por encima de 2.56 resulta evidente que la diferencia entre los estudiantes que opinan que el sueldo es lo más importante en la universidad pública y lo que opinan en la privada es muy significativa, por lo que se rechaza la hipótesis nula.
3. Como este valor calculado está muy por encima de 2.56 se rechaza la hipótesis nula y se concluye que la proporción de hombres a quienes les gusta el programa es diferente a la proporción de mujeres a las que les gusta.
5. Como este valor calculado está muy por debajo de –2.33 se rechaza la hipótesis nula y se concluye que la proporción de artículos defectuosos antes y después de realizar los cambios en la línea de producción son diferentes.

Ejercicios 10.4 Prueba para la diferencia entre dos varianzas

1. El valor empírico de F es mayor que 2.49, por lo que se rechaza la hipótesis nula y se concluye que la variabilidad de la vida útil de las luminarias que se fabrican en la línea de producción que está bajo sospecha es, efectivamente, mayor.
3. El valor empírico de F se encuentra entre 0.402 y 2.49 por lo cual se acepta la hipótesis nula y se concluye que las mediciones del nuevo supervisor son correctas.
5. El valor empírico de F se encuentra entre los valores 0.57 y 1.75 por lo que se acepta la hipótesis nula y se concluye que no se han presentado cambios en la varianza del tiempo de recorrido del maratón.

Capítulo 11 Pruebas estadísticas con la distribución ji cuadrada

Ejercicios 11.3 Pruebas de hipótesis para la varianza de una población

1. Como el valor calculado del estadístico es mayor que el valor crítico, se rechaza la hipótesis nula ya que la varianza no es la misma.
3. Como el valor calculado del estadístico es menor que el valor crítico, no se rechaza la hipótesis nula y podemos concluir que, efectivamente, el cálculo de la tarifa no es correcto.
5. Como el valor calculado del estadístico de prueba es menor que el valor crítico, no se rechaza la hipótesis nula y se concluye que la varianza del contenido de sodio en los productos de 1.5 litros está de acuerdo con lo establecido, es decir, es menor que 0.0020 miligramos cuadrados.

Ejercicios 11.5 Pruebas para una proporción con z y con χ^2

1. No es posible rechazar la hipótesis nula y se concluye que la proporción de trabajadores que utilizan este servicio es de, al menos, 50%.

3. No es posible rechazar la hipótesis nula y se concluye que, al menos, 60% de los clientes consume ese producto.
5. Es posible rechazar la hipótesis y se concluye que más de 80% de los trabajadores cuenta con tarjeta de crédito o débito.

Ejercicios 11.6 Prueba para la diferencia entre dos proporciones con z y con χ^2

1. No se puede rechazar la hipótesis nula y se concluye que ambas líneas de producción tienen la misma proporción de artículos que aprueban las evaluaciones.
3. No se puede rechazar la hipótesis nula y se concluye que la proporción de piezas vendidas entre las dos tiendas para uso doméstico es la misma.
5. No se puede rechazar la hipótesis nula y se concluye que la proporción de clientes de impuestos y de contabilidades que señalan quejas o inconformidades es la misma.

Ejercicios 11.8 Prueba para la diferencia entre n proporciones

1. Se acepta la hipótesis nula y se concluye que la proporción de hombres, mujeres y niños a los que les gusta la pasta dental es igual.
3. Se rechaza la hipótesis nula y se concluye que las proporciones de los expendios en las cuatro regiones que opinan que la renovación de los artículos es adecuada es significativamente diferente.
5. Se rechaza la hipótesis nula y se concluye que las tres proporciones de personas que recordaron no son iguales.

Ejercicios 11.9 Pruebas de bondad de ajuste a distribuciones teóricas

1. Se acepta la hipótesis nula y se concluye que los datos de la muestra se ajustan a una distribución normal.
3. Se rechaza la hipótesis nula y se concluye que los datos de la muestra no se ajustan a una distribución normal.
5. Se acepta la hipótesis nula y se concluye que los datos de esa muestra se ajustan a una distribución normal.
7. No se rechaza la hipótesis nula y se concluye que la distribución de accidentes diarios se ajusta a una distribución Poisson.
9. No se rechaza la hipótesis nula y se concluye que la distribución de pedidos tomados vía telefónica se ajusta a una distribución Poisson.
11. No se rechaza H_0 y se concluye que la distribución del número de refacciones robadas semanalmente se ajusta a una distribución binomial.
13. No se rechaza H_0 y se concluye que la distribución de empleados que faltan en cada sala se ajusta a una distribución binomial.

Ejercicios 11.10 Pruebas de bondad de ajuste entre distribuciones empíricas

1. Se acepta la hipótesis nula y se concluye que los dígitos tienen la misma probabilidad de ocurrir como terminación de la lotería ganadora.
3. Se acepta la hipótesis nula y se concluye que el número de retardos es igual todos los días de la semana.

Ejercicios 11.11 Pruebas sobre la independencia de dos variables

1. Se rechaza la hipótesis nula y se concluye que el tipo de defecto en las unidades producidas sí depende del turno en que son fabricadas.
3. Se rechaza la hipótesis nula y se concluye que el desempeño de los empleados en los cursos de capacitación depende de su desempeño en el trabajo.
5. Se rechaza la hipótesis nula y se concluye que la calidad de los limones depende de las regiones.

Capítulo 12 Análisis de varianza

Ejercicios 12.3 ANOVA con diseño completamente aleatorizado de un factor

1. *a*) H_0: $\mu_1 = \mu_2 = \mu_3$
   H_1: Por lo menos una de las medias poblacionales no es igual que las otras.
   c) Como el valor calculado de F = 11.81 es mayor que el valor crítico, 4.26, se rechaza la hipótesis nula y se concluye que el tiempo promedio que tardan los tres meseros para tomar la orden a una mesa no es igual.
   d) Se propone realizar pruebas entre pares de meseros para ver cuáles promedios difieren.
3. *a*) H_0: $\mu_1 = \mu_2 = \mu_3$
   H_1: Por lo menos una de las medias poblacionales no es igual que las otras.
   c) Como el valor calculado de F, 1.96, es menor que el valor crítico, 5.13, no es posible rechazar la hipótesis nula y se concluye que los ingresos promedio en las tres regiones son iguales.
5. *a*) H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$
   H_1: Por lo menos una de las medias poblacionales no es igual que las otras.
   c) Como el valor calculado de F, 3.95, es mayor que el crítico, 3.34, no se acepta la hipótesis nula y se concluye que existen diferencias entre los promedios de llenado de las cuatro máquinas.

Ejercicios 12.6 Pruebas entre pares de medias de tratamientos

De los ejercicios 1 a 5 de la sección anterior, sólo se rechazó la hipótesis nula en el 1 y en el 5. Se realizan en seguida las pruebas para pares de medias de tratamientos.

1. El administrador del restaurante puede concluir que no existe diferencia entre los meseros 1 y 3, pero que sí la hay entre ellos y el mesero 2.
5. Todos los pares de medias son significativamente diferentes entre sí, lo cual, a su vez, debe indicar al responsable del proceso de envasado que todas las máquinas están produciendo llenados diferentes y debe evaluar si esto es razonable, aceptable o no.

Ejercicios 12.7 ANOVA de dos factores

1. No existe diferencia en el tiempo de traslado al aeropuerto, entre las diferentes rutas ni entre los distintos conductores.
3. No existe diferencia en los tiempos de respuesta a los clientes, por turno ni por empleado.
5. No existe diferencia en el promedio de vehículos atendidos por hora entre las diferentes localizaciones y el número de competidores en las proximidades tampoco influye en el número de vehículos atendidos por hora.

Ejercicios 12.10 ANOVA de dos factores con interacción

1. *a*) Los promedios de los salarios para los tres puestos son iguales,
   b) los promedios de los salarios son iguales tanto para hombres como para mujeres y,
   c) no existe interacción entre los salarios por puesto y los salarios por sexo de los trabajadores.
3. Los promedios de calificaciones para los distintos horarios y para los distintos profesores son iguales, pero sí existe interacción entre horario de clase y maestro.
5. *a*) Todas las medias del factor A son iguales,
   b) todas las medias del factor B son iguales y
   c) no existe interacción entre los tres factores.

Capítulo 13 Análisis de regresión lineal simple

Ejercicios 13.2 Ecuación y recta de regresión

Los resultados pueden variar mucho y, por lo tanto, ser muy inexactos.

Ejercicios 13.3 Determinación de la ecuación de regresión

1. $\hat{y} = 0.73521 + 0.12709x$
3. $\hat{y} = 18.3849 - 1.0992x$
5. $\hat{y} = 16.1197 + 1.4773x$

Ejercicios 13.7 Inferencias estadísticas sobre la pendiente β_1

1. *a*) Se rechaza la hipótesis nula y se concluye que sí existe una relación entre las dos variables o, que β_1 es diferente de cero.
   b) Se estima, con un nivel de confianza de 99%, que el verdadero valor de $\hat{\beta}_1$ está entre 0.0099982 y 0.0154198.
3. *a*) Sí existe una relación entre las dos variables o, que $\hat{\beta}_1$ es diferente de cero.
   b) Se estima, con un nivel de confianza de 99%, que el verdadero valor de $\hat{\beta}_1$ está entre 0.933692 y 1.264708.
5. *a*) Se rechaza la hipótesis nula y se concluye que sí existe una relación entre las dos variables o, que β_1 es diferente de cero.
   b) Se estima, con un nivel de confianza de 99%, que el verdadero valor de $\hat{\beta}_1$ está entre 1.1842402 y 1.7703598.

Ejercicios 13.8 Estimación por intervalo y pronósticos de *y* para valores dados de *x*

1. *a*) 17.7172 a 21.8802
   b) 16.5127 a 23.0847
3. *a*) 5.3533 a 7.0349
   b) 3.8051 a 8.5831
5. *a*) 31 067.40 a 33 672.60
   b) 30 052.40 a 34 687.60

Ejercicios 13.9 Recapitulación del análisis de regresión lineal simple

Para los cinco ejercicios de la sección 13.3

a) Resuelto en el ejercicio 13.3
b) Resuelto en el ejercicio 13.3
c) Resuelto en el ejercicio 13.3
d) Resuelto en el ejercicio 13.4

1. *e*) Se rechaza la hipótesis nula y se concluye que sí existe una relación entre las dos variables o, que β_1 es diferente de cero.
3. *e*) Se rechaza la hipótesis nula y se concluye que sí existe una relación entre las dos variables o, que β_1 es diferente de cero.
5. *e*) Se rechaza la hipótesis nula y se concluye que sí existe una relación entre las dos variables o, que β_1 es diferente de cero.
   f) Resuelto en el ejercicio 13.3
   g) Resuelto en el ejercicio 13.8

Ejercicios 13.10 Análisis de correlación

1. *a*) 0.938
   b) Ver resultados en ejercicios 13.3

3. *a*) 0.953
   b) Ver resultados en ejercicios 13.3
5. *a*) 0.9675
   b) Ver resultados en ejercicios 13.3

Capítulo 14 Análisis de regresión lineal múltiple

Ejercicios 14.3 Multicolinealidad y las variables que mejor se relacionan con la variable dependiente: uso de la matriz de correlaciones

1. *a*) La variable B, pues es la independiente que tiene la correlación más alta con la dependiente y, además, no tiene correlación alta con ninguna de las otras variables independientes.
   b) La variable E, ya que es la que tiene la segunda correlación más alta con la variable dependiente y sólo tiene correlación alta con la variable D, la cual se eliminaría del modelo.
   c) Ver respuesta a *b*). Se eliminaría posiblemente también la variable C, ya que tiene una correlación muy pequeña con la variable dependiente.
3. *b*) Se incluye, en primer lugar, la variable de población urbana pues la correlación con la variable dependiente (0.5159) es la más alta. Así mismo la correlación con las demás variables independientes se encuentra en el rango de –0.70 y 0.70.
   c) La actividad económica terciaria pues no tiene correlación con las demás variables independientes, y la relación con la variable dependiente (0.4147) es la siguiente más alta a la de la población urbana.
   d) La correlación entre la actividad económica terciaria y la actividad económica primaria, es de –0.8289, la cual es demasiado alta y no está dentro del intervalo que se menciona como requisito. Como se debe eliminar una de las dos variables del modelo, se elimina la actividad económica primaria por ser la variable con menor relación con el PIB, nuestra variable dependiente.
   Por último, no se incluye en el modelo la actividad económica secundaria pues la correlación con el PIB es de sólo 0.1039.
   e) $y = -11\,018.79148 + 154.2650296x_1 + 122.1801185\,x_2$
5. *b*) En primer lugar, se incluye la potencia pues su correlación con la variable dependiente es mayor (0.8649), inclusión que elimina a las dos variables de millas por galón en ciudad y en carretera, ya que ambas tienen una alta correlación con aquélla.
   En segundo lugar, se incluye la variable *rpm* pues no guarda correlación con ninguna variable independiente.
   c) La correlación entre millas por galón en ciudad y millas por galón en carretera sobrepasa el 0.70, y de las dos la variable que tiene mayor correlación con la variable dependiente es millas por galón en carretera, con lo que se elimina del modelo las millas por galón en ciudad.
   Ahora bien, la variable de millas por galón en carretera tiene una alta correlación negativa con potencia, por lo tanto se elimina del modelo "millas por galón en carretera" pues la variable "potencia" tiene mayor correlación con la variable dependiente.
   d) $y = 48\,702.87295 + 117.1742938x_1 - 6.875479081x_2$

Ejercicios 14.4 Evaluación de la ecuación de regresión

1. *a*) El modelo en donde se incluyen las 4 variables independientes, no parece ser tan bueno, pues tan sólo 30.82% de la variación total se explica por la regresión.
   b) H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$
   H_1 = *Por lo menos una de las betas es distinta de cero*
   $\alpha = 0.01$
   $$F = \frac{CMR}{CME} = \frac{134\,080\,404.8}{30\,086\,729.53} = 4.45643261$$

La probabilidad de haber obtenido el valor anterior es tan sólo de 0.0045 (valor crítico de F), y siendo menor al nivel de significación, se concluye que sí existe relación lineal entre y y el conjunto de variables independientes, rechazándose la hipótesis nula.

   c) El modelo con sólo dos variables, quedó representado con la siguiente ecuación:
   $$y = -11\,018.79148 + 154.2650296x_1 + 122.1801185x_2$$
   x_1 *es la pobalción urbana*
   x_4 *es la actividad económica terciaria*

Se utilizará un nivel de significación de 1% ($\alpha = 0.01$)

Las hipótesis son:
H_0: $\beta_1 = 0$
H_0: $\beta_4 = 0$

Las hipótesis alternativas son:
H_1: $\beta_1 \neq 0$
H_1: $\beta_4 \neq 0$

Se puede afirmar, con un nivel de significación de 1%, que el coeficiente de la variable que es distinta de cero es el porcentaje de población urbana.

   d) Ahora el modelo es:
   $$\hat{y} = -6340.996413 + 193.3312708x_1$$

Se observa en la tabla que la probabilidad de obtener el valor de F es prácticamente cero (0.00028622). El coeficiente de determinación ajustado 0.249117876, indica que aproximadamente 24.92% de la variación en la variable dependiente queda explicada por el modelo, lo cual implica que no es un buen modelo.

3. *a*) Cerca de 99% de la variación total queda explicada por la regresión, en el modelo en el que se incluyen las 4 variables independientes.
   b)
   H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$
   H_1 = *Por lo menos una de las betas es distancia cero*
   $\alpha = 0.01$

$$F = \frac{CMR}{CME} = \frac{156\ 525\ 106.7}{6\ 696\ 806.416} = 23.37309712$$

La probabilidad de haber obtenido el valor anterior es de 0.1537 (valor crítico de F), y siendo mayor al nivel de significación, se rechaza la hipótesis nula y se concluye que no existe relación lineal entre y y el conjunto de variables independientes.

Aquí conviene revisar la matriz de correlaciones:

	Precio (USD)	Millas por galón en ciudad	Millas por galón en carretera	Potencia (hp)	rpm
Precio (USD)	1				
Millas por galón en ciudad	–0.714889	1			
Millas por galón en carretera	–0.84755	0.8774266	1		
Potencia (hp)	0.864933	–0.866599	–0.802137	1	
rpm	0.1397133	–0.770683	–0.50277	0.3830571	1

Se observa que todas las variables independientes tienen correlación alta con el precio, la variable dependiente. La potencia, que cuenta con la correlación más alta con el precio, tiene también correlación alta con las dos variables de millas, por lo que no se debe utilizar junto a éstas. La única otra variable con la que no tiene correlación alta es con la de revoluciones por minuto (rpm). Sin embargo, la correlación de rpm con el precio es muy baja, por lo que parece que la única variable útil es la de potencia. Corriendo el modelo con esta variable se obtiene un coeficiente de correlación de 0.748 y uno ajustado de 0.685.

Todo lo anterior lleva a concluir que esas cuatro variables independientes no constituyen información suficiente para construir un modelo de regresión lineal simple aceptable.

5. *a*) El modelo es el siguiente:

$$y = 1\ 240.256669 + 0.003438358x_1 + 282.2596949x_2$$

b) La regresión explica aproximadamente 70% de la variación total.

c)
H_0: $\beta_1 = \beta_2 = 0$
H_1 = *Por lo menos una de las betas es distancia cero*
$\alpha = 0.01$

$$F = \frac{CMR}{CME} = \frac{12\ 963\ 356.69}{254\ 574.6959} = 50.92162301$$

La probabilidad de haber obtenido el valor anterior, de acuerdo con los datos proporcionados por Excel, es de 0.00000000000355189, cifra mucho menor que nivel de significación de 0.01, por lo tanto se rechaza la hipótesis nula, es decir, por lo menos una de las betas es distinta de cero y sí existe relación lineal entre y y cuando menos una de las variables independientes.

d)
El modelo es:

$$y = 1\ 240.256669 + 0.003438358x_1 + 282.2596949x_2$$

Donde:
x_1 = *es el ingreso anual*
x_2 = *son los miembros de la familia*

Las hipótesis son:
H_0: $\beta_1 = 0$
H_0: $\beta_1 = 0$

Las hipótesis alternativas son:
H_0: $\beta_1 \neq 0$
H_1: $\beta_2 \neq 0$

El nivel de significación es de 0.01, por lo que se rechaza la hipótesis nula si el valor de probabilidad de t (5a. columna) es inferior. Tanto el ingreso anual como los miembros de la familia, implican el rechazo de nuestra hipótesis nula.

e) No es necesario modificar el modelo, pues ambas variables independientes son significativas.
f1) Se observa que la distribución tiene una forma aproximadamente normal.
f2) Poco más de la mitad de los valores son positivos, lo cual indica que la relación entre variables independientes con la variable dependiente es lineal.
f3) Los percentiles se ajustan a una recta con pendiente de 45°, así que los residuos del modelo, efectivamente, tienen una distribución aproximadamente normal.

Ejercicios 14.5 Uso del modelo de regresión lineal múltiple

1. 11 497.71343 a 13 500.83698.
3. 742.13487 a 56 185.62927 dólares.
5. $2 340.220834 a $4 682.068454.

Ejercicios 14.6 Variables independientes cualitativas

1. $\hat{y} = 5.019044285 + 0.078854865x_1 + 0.011570769x_2 + 0.984149028x_3$

Las primeras dos variables no requieren mayor análisis pues son numéricas; la variable x_3 es positiva. Por lo tanto, cuando equivale a 1, es decir, cuando el trabajador es hombre aumenta el número de faltas.

3. $\hat{y} = 86.44152109 + 16.80176431x_1 - 0.253470397x_2$

Cuando $x_1 = 1$ significa que las personas con capacitación logran hacer mayor número de repeticiones antes de cometer un error.

5. $\hat{y} = 17.46920904 + 2.695480226x_1 + 5.85x_2$

Cuando $x_2 = 1$, es decir, cuando el anaquel se ubica al principio del pasillo, las ventas del producto aumentan, esto lo sabemos porque la variable es positiva, así que cuando vale 0, para la ubicación en medio del pasillo, las ventas disminuyen.

Ejercicios 14.7 Regresión por pasos

1. El coeficiente de determinación es ahora 0.97202, lo cual es un valor más que aceptable. El modelo con estas 4 variables independientes es:

$$\hat{y} = 34.0696 + 3.38985x_1 - 0.72484x_2 + 0.36557x_3 - 0.18081x_5$$

Con la prueba F sobre este modelo se concluye que sí existe relación lineal entre y y el conjunto de variables independientes. Con las pruebas sobre los coeficientes del modelo se concluye que las cuatro variables son significativas para el modelo. Se puede afirmar, con un nivel de significación de 1%, que el coeficiente de cada una de las cuatro variables son distintos de cero.

Análisis de residuales

La distribución tiene una forma aproximadamente normal, salvo por la observación única del lado izquierdo. Esto permite pensar que se trata de un modelo aceptable que debería mejorarse eliminando esa observación extrema, aunque este análisis rebasa el alcance de este texto.

Los puntos del diagrama de dispersión de los residuos no siguen un patrón específico por lo que se concluye, de igual manera, que la relación entre la variable dependiente y la independiente es lineal.

Como los percentiles de la gráfica de probabilidad normal se ajustan aproximadamente a una recta con pendiente de 45°, se puede pensar que se distribuyen de forma normal aunque, al igual que con el histograma de los residuales, habría que abordar la cuestión de las primeras observaciones, las cuales distorsionan considerablemente la gráfica.

3. Como la superficie es la variable con la correlación más alta con el precio, se determina, en primer lugar, qué tanta variación de regresión queda explicada por esta variable. Se obtiene un coeficiente de determinación de 0.90520935.

   La única otra variable que no tiene correlación alta con la variable independiente que ya se introdujo es la de lugares de estacionamiento. Evaluando el modelo que incluyen estas dos variables independientes se obtiene un coeficiente de determinación múltiple de 0.90526918, lo cual indica que la contribución de la segunda variable que se introdujo al modelo, el número de lugares de estacionamiento, no contribuye prácticamente nada. Por esto, pareciera que el mejor modelo es el que se puede construir con una sola variable independiente, es decir, un modelo de regresión lineal simple.

   La prueba F sobre este modelo permite concluir que sí existe relación lineal entre y y la variable independiente y este resultado hace que sea innecesario llevar a cabo la prueba sobre el coeficiente de la única variable independiente utilizando la t de Student, ya que la prueba F ha demostrado que es diferente de cero.

Análisis de residuales

El histograma de los residuales tiene una forma que lejanamente se aproxima a una normal.

En el diagrama de dispersión de los residuos, los puntos del diagrama no siguen un patrón específico, por lo que se concluye que la relación entre la variable dependiente y la independiente es lineal.

Gráfica de probabilidad normal

Como los percentiles se ajustan aproximadamente a una recta con pendiente de 45°, se puede pensar que se distribuyen de forma normal aunque habría que abordar la cuestión de la primera observación, la cual distorsiona considerablemente la gráfica.

Capítulo 15 Números índice

Ejercicios 15.2 Números índice simples

1.

$$ISP_{pollo} = \frac{P_{2005}}{P_{2002}}(100) = \frac{34.10}{31.00}(100) = 110$$

$$ISP_{lenteja} = \frac{P_{2005}}{P_{2002}}(100) = \frac{7.3}{4.8}(100) = 152.08$$

$$ISP_{mantequilla} = \frac{P_{2005}}{P_{2002}}(100) = \frac{10.99}{7.75}(100) = 141.80$$

3.

$$ISP_{2005} = \frac{P_{2005}}{P_{2002}}(100) = \frac{14.72}{13.5}(100) = 109.03$$

$$ISP_{2004} = \frac{P_{2004}}{P_{2003}}(100) = \frac{13.72}{14.11}(100) = 97.23$$

5.

$$ISQ_{2003} = \frac{Q_{2003}}{Q_{2002}}(100) = \frac{0.87}{1.2}(100) = 72.5$$

$$ISQ_{2004} = \frac{Q_{2004}}{Q_{2002}}(100) = \frac{1.17}{1.2}(100) = 97.5$$

$$ISQ_{2005} = \frac{Q_{2005}}{Q_{2002}}(100) = \frac{1.28}{1.2}(100) = 106.66$$

7.

$$ISQ_{arroz} = \frac{Q_{2004}}{Q_{2002}}(100) = \frac{1.33}{0.86}(100) = 154.65$$

$$ISQ_{trigo} = \frac{Q_{2004}}{Q_{2002}}(100) = \frac{0.23}{0.11}(100) = 209.09$$

$$ISQ_{pasta} = \frac{Q_{2004}}{Q_{2002}}(100) = \frac{24}{16}(100) = 150$$

9.

$$ISV_{bolillo_2003} = \frac{P_{2003} \cdot Q_{2003}}{P_{2002} \cdot Q_{2002}}(100) = \frac{0.8 \cdot 27}{0.78 \cdot 24}(100)$$

$$= 115.38$$

$$ISV_{bolillo_2004} = \frac{P_{2004} \cdot Q_{2004}}{P_{2002} \cdot Q_{2002}}(100) = \frac{0.8 \cdot 31}{0.78 \cdot 24}(100)$$

$$= 132.47$$

$$ISV_{bolillo_2005} = \frac{P_{2005} \cdot Q_{2005}}{P_{2002} \cdot Q_{2002}}(100) = \frac{0.9 \cdot 34}{0.78 \cdot 24}(100)$$

$$= 163.46$$

$$ISV_{atún_{2003}} = \frac{P_{2003} \cdot Q_{2003}}{P_{2002} \cdot Q_{2002}}(100) = \frac{7.9 \cdot 6.02}{ND}(100) = ND$$

$$ISV_{atún_{2004}} = \frac{P_{2004} \cdot Q_{2004}}{P_{2002} \cdot Q_{2002}}(100) = \frac{6.14 \cdot 5.3}{ND}(100) = ND$$

$$ISV_{atún_{2005}} = \frac{P_{2005} \cdot Q_{2005}}{P_{2002} \cdot Q_{2002}}(100) = \frac{5.15 \cdot 4}{ND}(100) = ND$$

Ejercicios 15.3 y 15.4 Números índice agregados

3. *a*)

$$IAP_{2011} = \frac{\sum P_{2011}}{\sum P_{2008}}(100) = \frac{7.05 + 7 + 7.25}{6.5 + 6.65 + 6.55}(100)$$

$$= \frac{21.3}{19.7}(100) = 108.12$$

b)

$$IAV_{2011} = \frac{\sum P_{2011} \times Q_{2011}}{\sum P_{2008} \times Q_{2008}}(100)$$

$$= \frac{7.05(200) + 7(250) + 7.25(170)}{6.5(200) + 6.65(250) + 6.55(170)}(100)$$

$$= \frac{4\ 392.5}{4\ 076}(100) = 107.76$$

5. *a*)

$$IAP_{Septiembre} = \frac{\sum P_{Septiembre}}{\sum P_{Agosto}}(100)$$

$$= \frac{79.9 + 26.8 + 60 + 32.6}{67.5 + 18 + 54.7 + 37.4}(100)$$

$$= \frac{199.3}{177.6}(100) = 112.21$$

b)

$$IPP = \frac{\sum P_{Septiembre} \times Q}{\sum P_{Agosto} \times Q}(100)$$

$$= \frac{79.9(30) + 26.8(40) + 60(20) + 32.6(10)}{67.5(30) + 18(40) + 54.7(20) + 37.4(10)}(100)$$

$$= \frac{4\ 995}{4\ 213}(100) = 118.56$$

7. *a*)

$$IPP_{Agosto} = \frac{\sum P_{Agosto} \times Q_{Agosto}}{\sum P_{Mayo} \times Q_{Mayo}}(100)$$

$$= \frac{239(502) + 430(176) + 600(100)}{223(438) + 354(153) + 543(200)}(100)$$

$$= \frac{255\ 658}{260\ 436}(100) = 98.16$$

b)

$$IVL_{Agosto} = \frac{\sum P_{Agosto} \times Q_{Mayo}}{\sum P_{Mayo} \times Q_{Mayo}}(100)$$

$$= \frac{239(438) + 430(153) + 600(200)}{223(438) + 354(153) + 543(200)}(100)$$

$$= \frac{290\ 472}{260\ 436}(100) = 111.53$$

c)

$$IVP_{Agosto} = \frac{\sum P_{Agosto} \times Q_{Agosto}}{\sum P_{Mayo} \times Q_{Agosto}}(100)$$

$$= \frac{239(502) + 430(176) + 600(100)}{223(502) + 354(176) + 543(100)}(100)$$

$$= \frac{255\ 658}{228\ 550}(100) = 111.86$$

Capítulo 16 Análisis de series de tiempo

Ejercicios 16.3 Tendencia secular

7. La ecuación de regresión lineal simple:

$$\hat{y} = 26.21397058 - 0.559068627x$$

La tasa de natalidad para 2015: 14.4735294

16.4 Variaciones estacionales

1. *a*)

Año	Mes	Cociente entre datos y promedio móvil				Promedio	Promedio ajustado a 12 (índices estacionales)
		2007	2008	2009	2010		
2007	Ene.				1.01415759	1.01415759	1.01709769
	Feb.			1.074397467		1.074397467	1.077512205
	Mar.			1.039389294		1.039389294	1.042402542
	Abr.			0.982513713		0.982513713	0.985362075
	Mayo				1.010227524	1.010227524	1.01315623
	Jun.				1.015581009	1.015581009	1.018525235
	Jul.			0.949643103		0.949643103	0.952396172
	Ago.			0.948504609		0.948504609	0.951254377
	Sep.		0.95502196			0.95502196	0.957790622
	Oct.	0.973119707				0.973119707	0.975940836
	Nov.	1.015089955				1.015089955	1.018032758
	Dic.	0.98766596				0.98766596	0.990529259
					Sumas	11.96531189	12

b)

					Promedio ajustado a 12 (índices estacionales)	Precios desestacionalizados			
	2007	2008	2009	2010		2007	2008	2009	2010
Ene.	631.17	889.6	858.69	1 117.96	1.01709769	620.5598601	874.6455813	844.2551868	1 099.166787
Feb.	664.75	922.3	943	1 095.41	1.077512205	616.9303667	855.9531812	875.1641005	1 016.610294
Mar.	654.9	968.43	924.27	1 113.34	1.042402542	628.2601718	929.0364914	886.6728188	1 068.051885
Abr.	679.37	909.71	890.2	1 148.69	0.985362075	689.4622972	923.2240847	903.4242563	1 165.754223
Mayo	667.31	888.66	928.65	1 205.43	1.01315623	658.6447185	877.1204021	916.5911163	1 189.77702
Jun.	655.66	889.49	945.67	1 232.92	1.018525235	643.7346639	873.3116954	928.4698771	1 210.49529
Jul.	665.38	939.77	934.23	1 192.97	0.952396172	698.6378357	986.7427317	980.9258247	1 252.598483
Ago.	665.41	839.03	949.38	1 215.81	0.951254377	699.5079509	882.0248509	998.0295733	1 278.112384
Sep.	712.65	829.93	996.59	1 270.98	0.957790622	744.0561471	866.504621	1 040.509248	1 326.991485
Oct.	754.6	806.62	1 043.16	1 342.02	0.975940836	773.2026087	826.5050202	1 068.87627	1 375.10385
Nov.	806.25	760.86	1 127.04	1 369.89	1.018032758	791.9686218	747.3826302	1 107.07636	1 345.624676
Dic.	803.2	816.09	1 134.72	1 390.55	0.990529259	810.879631	823.8928761	1 145.569391	1 403.845457

c) y *d*)

Mes		2007
Ene. 07		631.17
Feb. 07		664.75
Mar. 07		654.9
Abr. 07		679.37
Mayo 07		667.31
Jun. 07		655.66
Jul. 07		665.38
Ago. 07		665.41
Sep. 07		712.65
Oct. 07		754.6
Nov. 07		806.25
Dic. 07		803.2
Ene. 08		889.6
Feb. 08		922.3
Mar. 08		968.43
Abr. 08		909.71
Mayo 08		888.66
Jun. 08		889.49
Jul. 08		939.77
Ago. 08		839.03
Sep. 08		829.93
Oct. 08		806.62
Nov. 08		760.86
Dic. 08		816.09
Ene. 09	1	858.69
Feb. 09	2	943
Mar. 09	3	924.27
Abr. 09	4	890.2
Mayo 09	5	928.65
Jun. 09	6	945.67
Jul. 09	7	934.23
Ago. 09	8	949.38
Sep. 09	9	996.59
Oct. 09	10	1 043.16
Nov. 09	11	1 127.04
Dic. 09	12	1 134.72
Ene. 10	13	1 117.96
Feb. 10	14	1 095.41
Mar. 10	15	1 113.34
Abr. 10	16	1 148.69
Mayo 10	17	1 205.43
Jun. 10	18	1 232.92
Jul. 10	19	1 192.97
Ago. 10	20	1 215.81
Sep. 10	21	1 270.98
Oct. 10	22	1 342.02
Nov. 10	23	1 369.89
Dic. 10	24	1 390.55
Ene. 11	25	1 370.68645
Feb. 11	26	1 392.43613
Mar. 11	27	1 414.18581
Abr. 11	28	1 435.9355
Mayo 11	29	1 457.68518
Jun. 11	30	1 479.43486
Jul. 11	31	1 501.18454
Ago. 11	32	1 522.93423
Sep. 11	33	1 544.68391
Oct. 11	34	1 566.43359
Nov. 11	35	1 588.18328
Dic. 11	36	1 609.93296

Con los datos anteriores, se determinó que la ecuación de regresión es:

$$y = 826.9443841 + 21.74968261x$$

e) Precios desestacionalizados de 2011

Índices estacionales	Precios desestacionalizados
1.01709769	1 394.122021
1.077512205	1 500.366926
1.042402542	1 474.150888
0.985362075	1 414.916381
1.01315623	1 476.862821
1.018525235	1 506.84174
0.952396172	1 429.722414
0.951254377	1 448.69785
0.957790622	1 479.483763
0.975940836	1 528.74651
1.018032758	1 616.8226
0.990529259	1 594.6857

f) Los precios reales de acuerdo con el sitio citado, de enero a octubre de 2011 son los siguientes:

Ene. 2011	1 356.40
Feb. 2011	1 372.73
Mar. 2011	1 424.00
Abr. 2011	1 479.76
Mayo 2011	1 512.60
Jun. 2011	1 528.66
Jul. 2011	1 572.21
Ago. 2011	1 757.21
Sep. 2011	1 770.95
Oct. 2011	1 665.21

3. *a*) Resumen de cocientes después de haber eliminado máximos y mínimos por renglón y cálculo de índices estacionales.

Año	**Mes**	**Cociente**				**Promedio**	**Índices estacionales**
		2007	2008	2009	2010		
2007	**Ene.**			0.926243568		0.926243568	0.927175077
	Feb.			0.918367347		0.918367347	0.919290935
	Mar.			0.910623946		0.910623946	0.911539747
	Abr.			0.844221106		0.844221106	0.845070126
	Mayo			0.84		0.84	0.840844775
	Jun.		0.830523514			0.830523514	0.831358759
	Jul.			1.10982659		1.10982659	1.110942726
	Ago.	1.095890411				1.095890411	1.096992532
	Sep.	1.07946027				1.07946027	1.080545867
	Oct.			1.147996729		1.147996729	1.149151252
	Nov.			1.144254279		1.144254279	1.145405038
	Dic.			1.140536149		1.140536149	1.14168317
					Sumas	**11.98794391**	**12**

b)

Año	**Mes**	**Ventas**				**Índices estacionales**	**Ventas desestacionalizadas**			
		2007	2008	2009	2010		2007	2008	2009	2010
2007	**Ene.**	24	29	30	31.3333333	0.927175077	25.88507888	31.27780364	32.35634859	33.79440853
	Feb.	24	29	30	31.3333333	0.919290935	26.10707784	31.54605239	32.6338473	34.08424051
	Mar.	24	29	30	31.3333333	0.911539747	26.3290768	31.81430113	32.911346	34.37407249
	Abr.	20.66666667	26	28	29.3333333	0.845070126	24.45556414	30.76667746	33.13334496	34.71112329
	Mayo	20.66666667	26	28	29.3333333	0.840844775	24.57845642	30.92128388	33.29984418	34.88555105
	Jun.	20.66666667	26	28	29.3333333	0.831358759	24.85890291	31.27410366	33.67980394	35.28360413
	Jul.	30	34	37.33333333	40	1.110942726	27.00409239	30.60463804	33.60509275	36.00545652
	Ago.	30	34	37.33333333	40	1.096992532	27.34749703	30.99382997	34.03244075	36.46332938
	Sep.	30	34	37.33333333	40	1.080545867	27.76374508	31.46557776	34.55043833	37.01832678
	Oct.	32.33333333	37	39	42	1.149151252	28.13670808	32.19767626	33.93809119	36.54871359
	Nov.	32.33333333	37	39	42	1.145405038	28.22873329	32.30298346	34.04909067	36.66825149
	Dic.	32.33333333	37	39	42	1.14168317	28.3207585	32.40829066	34.16009015	36.7877894

c) y *d*)

Las ventas para 2011 se calcularon con la siguiente ecuación de regresión, misma que se obtuvo de los resultados de "Regresión" de Excel:

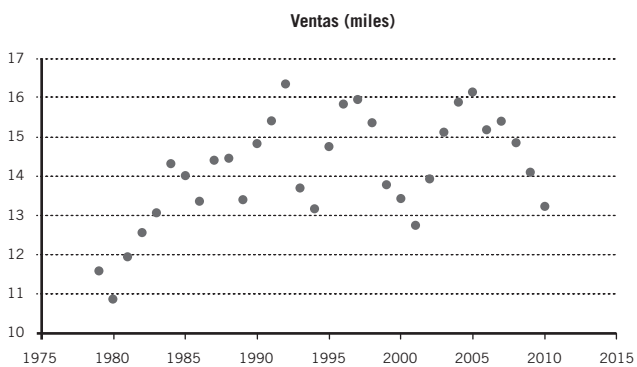
$$y = 24.6143617 + 0.296352584x$$

e)

Fecha	Pronóstico de ventas	Índices estacionales	Ventas desestacionalizadas
Ene. 11	39.13563832	0.927175077	36.28558847
Feb. 11	39.4319909	0.919290935	36.24947179
Mar. 11	39.72834348	0.911539747	36.21396416
Abr. 11	40.02469607	0.845070126	33.82367495
Mayo 11	40.32104865	0.840844775	33.9037431
Jun. 11	40.61740124	0.831358759	33.76763228
Jul. 11	40.91375382	1.110942726	45.45283718
Ago. 11	41.2101064	1.096992532	45.20717895
Sep. 11	41.50645899	1.080545867	44.84963271
Oct. 11	41.80281157	1.149151252	48.03775327
Nov. 11	42.09916416	1.145405038	48.22059472
Dic. 11	42.39551674	1.14168317	48.40224793

16.5 Variaciones cíclicas

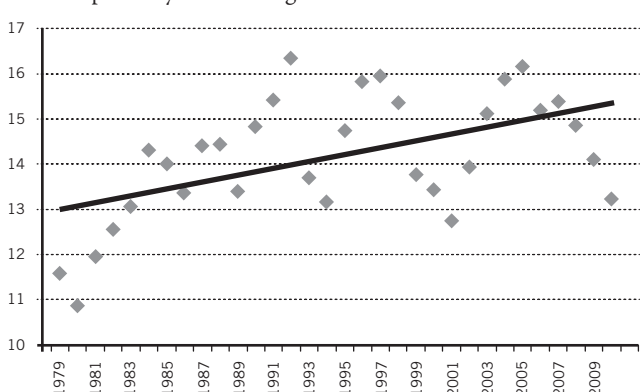
1. *a*)



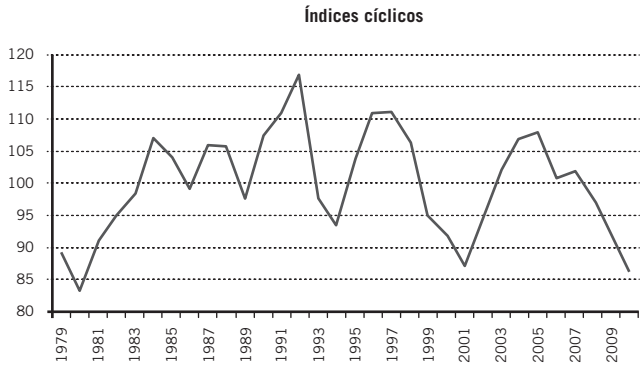
b) La ecuación de regresión es la siguiente:

$$\hat{y} = 12.90453629 + 0.075975073x$$

Nube de puntos y recta de regresión:



Gráfica de índices cíclicos:



No parece haber variaciones cíclicas

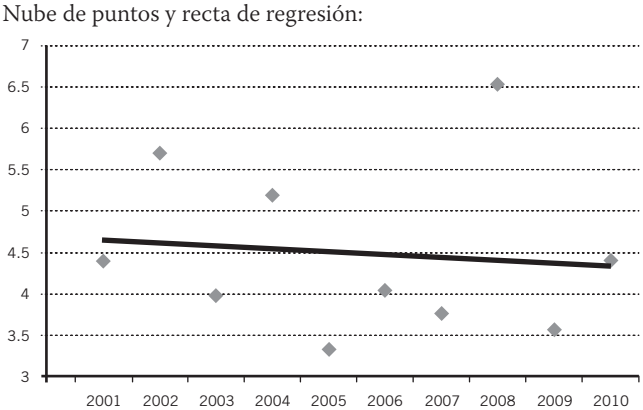
3. *a*)
b)

Año	Inflación %	Año x	$(X-\bar{X})^2$	$(X-\bar{X})(Y-\bar{Y})$	$\hat{y}$	Índices cíclicos
2001	4.4	1	20.25	0.4095	4.647272727	94.67918623
2002	5.7	2	12.25	–4.2315	4.612545454	123.576018
2003	3.98	3	6.25	1.2775	4.577818181	86.94098024
2004	5.19	4	2.25	–1.0485	4.543090908	114.2394045
2005	3.33	5	0.25	0.5805	4.508363635	73.86271982
2006	4.05	6	0.25	–0.2205	4.473636362	90.53038004
2007	3.76	7	2.25	–1.0965	4.438909089	84.70549688
2008	6.53	8	6.25	5.0975	4.404181816	148.2681749
2009	3.57	9	12.25	–3.2235	4.369454543	81.70356196
2010	4.4	10	20.25	–0.4095	4.33472727	101.5058094
Promedios	4.491	5.5				
		Sumas	82.5	–2.865		

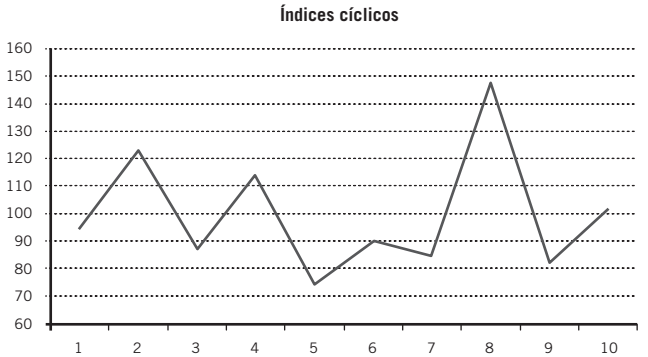
La ecuación de regresión es la siguiente:

$$\hat{y} = 4.682 - 0.034727273x$$

Nube de puntos y recta de regresión:



Gráfica de índices cíclicos:



Capítulo 17 Estadística no paramétrica

1. La probabilidad de encontrar 28 rachas en las condiciones dadas es 0.1024, o 10.24% y, como esta probabilidad es mayor que el nivel de significación, no se rechaza la hipótesis nula y se concluye que el paso de automóviles con placas del D.F. y con placas foráneas por el punto de observación es aleatorio.

Ejercicios 17.3 Prueba de los signos

1. La probabilidad de tener estos 5 trabajadores que superan la satisfacción mediana en el trabajo, es de 0.3633 y, como esta probabilidad es superior al 0.01 del nivel de significación, no se rechaza la hipótesis nula y se concluye, por lo tanto, que la mediana de la satisfacción de los empleados con el trabajo sigue siendo de 70.
3. La probabilidad de tener estas 7 observaciones, o más, es de 0.2744 y, como esta probabilidad es superior al 0.05 del nivel de significación, no se rechaza la hipótesis nula y se concluye, por lo tanto, que la mediana de los salarios sectoriales medianos es de cuando menos $6 000.
5. La probabilidad de tener estas 4 observaciones, o más, es de 0.5001 y, como esta probabilidad es muy superior al 0.005 del nivel de significación, no se rechaza la hipótesis nula y se concluye, por lo tanto, que el contenido de cereal de esa presentación es de, efectivamente 335 gramos.
7. Como este valor empírico de z de -1.33 es mayor que el valor crítico de -1.645, no se rechaza la hipótesis nula y se concluye que la mediana de los ingresos de las familias de ese municipio es de cuando menos $4 000.
9. Como este valor empírico de z de 0.73 es mayor que el valor crítico de -2.33, no se rechaza la hipótesis nula y se concluye que la mediana del tiempo de espera de los pacientes es mayor que 25 minutos.
11. Como este valor empírico de z de -1.97 es mayor que -1.96, se rechaza la hipótesis nula y se concluye que el contenido de las botellas de ese limpiador no es 1.05 litros en promedio.
13. Como esta probabilidad de 0.1719 es superior al 0.05 del nivel de significación, no rechaza la hipótesis nula y se concluye, por lo tanto, que no hay diferencia entre las tasaciones de los dos valuadores.
15. Como esta probabilidad de 0.0625 es superior al 0.025 del nivel de significación, no rechaza la hipótesis nula y se concluye que no hay diferencia entre el peso de las personas antes y después de la dieta.
17. Como este valor empírico de z de 1.37 es menor que valor crítico de 1.645, no se rechaza la hipótesis nula y se concluye que no aumentó el flujo de clientes con la reestructuración.
19. Como este valor empírico de z de 1.47 es menor que valor crítico de 1.645, no se rechaza la hipótesis nula y se concluye que el tratamiento de alcohólicos no mejoró su patrón de conducta.
21. Como este valor empírico de z de 0.79 es menor que valor crítico de 1.645, no se rechaza la hipótesis nula y se concluye que la presión sistólica no bajó con el ejercicio.

Ejercicios 17.4 Prueba de rangos con signo de Wilcoxon

1. Como el estadístico de prueba crítico es $z = \pm 1.96$ y el valor calculado cae dentro del intervalo, no se rechaza la hipótesis nula y se concluye que la mediana de las reclamaciones no ha cambiado, misma conclusión a la que se llegó en el ejemplo 17.7.
3. Como el estadístico de prueba crítico es $z = 2.33$, y como el valor calculado es menor, no es posible rechazar la hipótesis nula y se concluye que la preferencia por la marca 1 aún es menor o igual que la preferencia por la marca 2. Es la misma conclusión a la que se llegó en el ejemplo 17.9.

Prueba de rangos con signo de Wilcoxon para una muestra pequeña

5(1). El valor observado del estadístico T es 6, atendiendo al sentido de la hipótesis nula. Para $n = 8$, una prueba de un extremo y $\alpha = 0.01$, el valor crítico de T es 2, y siendo menor que el valor observado, no es posible rechazar la hipótesis nula, por lo que el nivel de satisfacción en el trabajo no ha aumentado.

5(3). El valor observado del estadístico T es 13, atendiendo al sentido de la hipótesis nula. Para $n = 11$, una prueba de un extremo y $\alpha = 0.05$, el valor crítico de T es 14 y siendo mayor que el valor observado, se rechaza la hipótesis nula, por lo tanto, la afirmación del funcionario es correcta, pues los salarios de los profesionales mencionados superan los $6 000.

5(5). El valor observado del estadístico T es 19, por ser ésta la suma menor de rangos de signo, para una prueba de dos extremos. Para $n = 9$, una prueba de dos extremos y $\alpha = 0.01$, el valor crítico de T es 2 y siendo menor al valor observado, no es posible rechazar la hipótesis nula, por lo que efectivamente la presentación del cereal tiene una mediana de 335 gramos.

6(7). Como el estadístico de prueba crítico es $z = 1.645$, y el valor calculado es menor, 0.9655, no se rechaza la hipótesis nula y se concluye que, efectivamente, los ingresos familiares están por debajo de $4 000, por lo que se puede sostener la petición de fondos del funcionario.

6(9). Como el estadístico de prueba crítico es $z = -1.645$, y el valor calculado es mayor, 24.8495, se rechaza la hipótesis nula, así que se puede afirmar que el tiempo de espera de los pacientes no rebasa los 25 minutos.

6(11). Como el estadístico de prueba crítico es z = ± 1.96, y el valor calculado, 1.74, se encuentra dentro de la región de aceptación, no es posible rechazar la hipótesis nula y se concluye que la afirmación del fabricante es verdadera.

7(13). El valor empírico del estadístico T es igual que 8.5, que es la menor de las dos sumas de rangos con signo. Conforme los valores de la tabla, para $n = 10$, una prueba de dos extremos y $\alpha = 0.05$, el valor crítico de T es 8, y como el valor observado de T es mayor, no se rechaza la hipótesis nula y por lo tanto no existe diferencia entre los criterios de los valuadores.

7(15). El valor empírico del estadístico T es igual que 1, que es la menor de las dos sumas de rangos con signo. Conforme los valores de la tabla, para $n = 7$ (porque se eliminó una observación), una prueba de dos extremos y $\alpha = 0.05$, el valor crítico de T es 2, y como el valor observado de T es menor, se rechaza la hipótesis nula y por lo tanto sí existe diferencia entre el peso antes y después de la dieta.

7(17). Como el estadístico de prueba crítico es $z = 1.645$, y el valor calculado, –3.2, es menor, no se rechaza la hipótesis nula, así que se puede afirmar que la reestructuración de las cajas no mejoró la afluencia de clientes.

7(19). Como el estadístico de prueba crítico es $z = 1.645$, y el valor calculado es menor, 1.6, no es posible rechazar la hipótesis nula, por lo que la participación de los psiquiatras no mejoró la conducta de los alcohólicos.

7(21). Como el estadístico de prueba crítico es $z = 1.645$, y el valor calculado es menor, 0.9784, no se rechaza la hipótesis nula, por lo que sí disminuyó la presión sanguínea de las personas que hicieron ejercicio.

Ejercicios 17.5 Prueba de Mann-Whitney para 2 muestras independientes

1. Por lo que el estadístico de prueba calculado o empírico es $U = 27$. Como el valor crítico que se determinó antes es 27 y la U calculada es mayor que 27, no es posible rechazar la hipótesis nula y se concluye que no existe diferencia entre los gastos en alimentación de las familias de la colonia 1 con los gastos de las familias de la colonia 2. Con la aproximación normal: para un nivel de significación de 0.05 y una prueba de dos colas, el valor crítico de z es 1.96, ya que $P(-1.96 \leq z \leq 1.96)$. Como el valor calculado de la z, 0.4535 está entre –1.96 y 1.96, no se rechaza la hipótesis nula y se concluye, igualmente, que no existe diferencia entre los gastos en alimentación de las familias de la colonia 1 con los gastos de las familias de la colonia 2.
3. El estadístico de prueba calculado o empírico es $U_1 = 4$, pues se trata de una prueba de un extremo. Como el valor crítico que se determinó antes es 16 y la U_1 calculada es menor que la U crítica (16) se rechaza la hipótesis nula, por lo que se concluye que los pronósticos de los economistas del sector público tienden a ser mayores que aquellos de los economistas del sector privado.
5. Por lo que el estadístico de prueba calculado o empírico es $U = 29$. Como el valor crítico que se determinó antes es 16 y la U calculada es mayor que 16, no es posible rechazar la hipótesis nula y se concluye que no existe diferencia entre las casas de cambio del aeropuerto de la Ciudad de México y las casas de cambio en aeropuertos de ciudades fronterizas mexicanas.

Ejercicios 17.6 Prueba de Kruskal-Wallis para más de 2 medias

1. Como el valor empírico (calculado) del estadístico de prueba, 7.791, es mayor que el valor crítico, se rechaza la hipótesis nula y se concluye que cuando menos una de las medias no es igual que las otras.
3. Como el valor empírico (calculado) del estadístico de prueba, 16.794, es mayor que el valor crítico, se rechaza la hipótesis nula y se concluye que cuando menos una de las medias no es igual que las otras.
5. Como el valor empírico (calculado) del estadístico de prueba, 13.391, es mayor que el valor crítico, se rechaza la hipótesis nula y se concluye que cuando menos una de las medias no es igual que las otras.

Ejercicios 17.7 Prueba de Friedman para diseños en bloques aleatorizados

1. Como el valor empírico (calculado) del estadístico de prueba, 8.106, es mayor de 5.991, se rechaza la hipótesis nula y se concluye que cuando menos una de las medias no es igual que las otras.
3. Como el valor empírico (calculado) del estadístico de prueba, 3.74, es menor de 9.21, no se rechaza la hipótesis nula y se concluye que todas las medias son iguales.
5. Como el valor empírico (calculado) del estadístico de prueba, 13.92, es mayor de 11.143, se rechaza la hipótesis nula y se concluye que cuando menos una de las medias no es igual que las otras.

Ejercicios 17.8 Coeficiente de correlación de rangos de Spearman

1. Como el valor calculado del estadístico de prueba, –3.039 es menor que 2.306, se rechaza la hipótesis nula y se concluye que el coeficiente de correlación por rangos de Spearman no es igual que cero y, por lo tanto, se concluye también que sí existe relación entre la cantidad de habitantes y la calidad del aire en las ciudades y que, dado el valor negativo de r_r, se trata de una relación inversa.
3. Como el valor calculado del estadístico de prueba, 3.43 es mayor que 2.131, se rechaza la hipótesis nula y se concluye

que el coeficiente de correlación por rangos de Spearman no es igual que cero y, por lo tanto, se concluye también que sí existe relación entre las calificaciones de los estudiantes a la mitad y al final del curso. Además, como el coeficiente r_r es positivo, se concluye que la relación entre esas dos variables es directa.

5. Como el valor calculado del estadístico de prueba, 3.507 es mayor que 3.106, se rechaza la hipótesis nula y se concluye que el coeficiente de correlación por rangos de Spearman no es igual que cero y, por lo tanto, se concluye también que sí existe relación entre los dos tipos de evaluaciones de candidatos y, además, que esa relación es directa.

Índice analítico

a

Ajuste de una
  función exponencial con mínimos cuadrados, 482
  parábola con mínimos cuadrados, 486
  recta con mínimos cuadrados, 481
Análisis de correlación, 399
Análisis de la tendencia secular, 479
Análisis de regresión
  correlación, 2
  lineal simple, 375, 393
  múltiple, 374
  simple, 374
  y suma de los cuadrados, 387
Análisis de residuales, 425
Análisis de series de tiempo, 2
Análisis de varianza, 419 *también véase* ANOVA
  en Excel, 366, 361
Análisis multivariado, 3
ANOVA, 345
  de dos factores, 347
Aproximación de la distribución binomial, 197
Área(s)
  entre, dos valores negativos de *z*, 183
  entre, dos valores positivos de *z*, 182
  bajo la curva de la distribución χ^2, 317
  bajo la curva normal, 179
Ausencia de sesgo, 241
Autocorrelación, 403
Axioma, 124

b

Banco de México, 29, 464
Bolsa Mexicana de Valores, 113, 401
Bondad de ajuste, 269, 326

c

Cálculo de la inflación, 465
Cambio de periodo, 465
Campana de Gauss, 31, 76
Canasta básica, 463
Censo, 1
Ciclo de negocios, 477
Clave de Yahoo! para 34 índices, 469
Coeficiente
  de confianza, 241
  de correlación momento-producto de Pearson, 402
  de correlación por rangos de Spearman, 541
  de correlación serial de orden, 1, 404
  de determinación, 399
  de determinación múltiple, 418
  de regresión parcial, 422
  de variación, 74
Coeficiente de curtosis, 83
  en una serie de clases y frecuencias, 86
  en una serie de datos y frecuencias, 85
  en una serie simple, 84
Coeficiente de sesgo, 44
  en una serie de clases y frecuencias, 86
  en una serie de datos y frecuencias, 85
  en una serie simple, 84
Combinaciones, 118
Comisión Nacional de los Salarios Mínimos *véase* Conasami
Complemento, 110
Conasami, 12, 17
Conclusiones, 5
Conglomerados, 215
Consistencia, 242
Conversión de series simples, 17
Cuartiles, 57, 58
Curtosis, 82
Curva característica operativa, 273

d

Datos, 1
  continuos, 8
  cualitativos en escala nominal, 20
  de fuentes externas, 5
  de fuentes internas, 5
  de fuentes primarias, 5
  de fuentes secundarias, 5
  discretos o discontinuos, 8
  pareados, 293
Deciles, 57, 59

Deflación de series de tiempo, 466
Desviación estándar, 67, 71, 153
en una serie de clases y frecuencias, 73
en una serie de datos y frecuencias, 72
de la distribución hipergeométrica, 170
de la distribución multinomial, 170
de regresión, 388
Desviación intercuartílica, 71
Desviación media, 69, 70
Determinación de *z* a partir del área, 184
Diagrama de dispersión, 374
Diagramas de Pareto, 32
Diagramas de Venn-Euler, 109
Diferencias en relación con la distribución binomial, 148
Discretos, 8
Diseño de la muestra, 215
Distribución binomial, 151, 154
fórmula, 170
Distribución de Poisson, 148, 158, 161, 204
fórmula, 170
media y varianza, 162
Distribución de probabilidad, 149
para una variable aleatoria discreta, 147
teórica, 156
Distribución *F* de Fisher, 204, 348
Distribución hipergeométrica, 153, 163
fórmula, 170
aritmética y varianza, 165
Distribución χ^2 (ji cuadrada), 160, 292, 316
Distribución leptocúrtica, 83
Distribución mesocúrtica, 83
Distribución muestral, 216
de la media, 216
Distribución multinomial, 148, 168
fórmula, 170
medias y varianzas, 169
Distribución normal, 31, 76, 178
estándar, 76, 316
Distribución normal (*z*), 250
Distribución platicúrtica, 83
Distribución Poisson, 329
Distribución *t* de Student, 204, 249, 250, 512
Distribuciones
de frecuencias, 12
de probabilidad binomial, 148
discretas de probabilidad, 148

e

Ecuación de regresión
evaluación, 418, 429
de mínimos cuadrados, 381
lineal, 380
Ecuaciones normales, 378, 394
Ecuación normal I, 378
Ecuación normal II, 379
Eficiencia, 242
Elaboración de tablas y gráficas, 1
Elementos de la población, 4
Eliminación posterior, 437
Error de muestreo, 241
Errores que no se deben al muestreo, 241
Error estándar
de la diferencia entre dos medias, 294
de la media, 218
Error típico, 432
Error tipo I y II, 270
Escala
de intervalo, 7
nominal, 6
Escalas
de medición, 511
ordinales, 7
Espacio muestral, 113, 124
Espacio muestral del experimento, 156
Estadística, 1
Estadística descriptiva, 3
opción de Excel, 90
Estadística inferencial, 2, 3
Estadísticas de población, 1
Estadísticas de ventas, 1
Estadístico, 4, 240, 244
Estadístico muestral *véase* Estadístico
Estadísticos muestrales, 210
Estadístico *T* de Wilcoxon, 527
Estimación de parámetros, 211
Estimaciones de la media
intervalos de confianza, 432
Estimación por intervalo, 240
ecuación de regresión, 391
Estimación por punto, 211, 240
Estimador, 210, 229
Estimadores insesgados, 230, 233, 241
Estimador sesgado, 230, 241
Estructura de cálculo, 469
Evento, 108
compuesto, 114
simple, 114
Eventos independientes, 133
Eventos mutuamente excluyentes, 133
Excel, 9, 29, 30
coeficiente de correlación, 400, 402
función \fCOMBINAT\C, 516
números índices, 473
prueba de rachas de Wald-Wolfowitz, 516
prueba de Wilcoxon, 528
regresión, 419, 425, 485
Experimento aleatorio, 112, 115

f

Factor de corrección
por discontinuidad, 523
por población finita, 220
Factor de suavización, 479
F de Fisher, 345, 438
Fórmula
de las combinaciones, 231

del error estándar, 219
Frecuencia(s)
absolutas, 20
de aparición, 16
de la clase anterior, 57
de la clase modal, 57
de la clase siguiente, 57
esperada de éxitos, 319
relativa, 121, 122, 149
relativas, 20
Función
\fCOMBINAT\C, 516
\lDistr.T.Inv\D, 250
\fDistr.Binom, 330
COEFICIENTE.ASIMETRIA, 84
CUARTIL, 89
CURTOSIS, 85
de densidad de probabilidad, 179
DESVPROM, 69
Distr.Binom, 330
DISTR.BINOM, 154
Distr.Chi, 338
DISTR.F, 307
DISTR.F.INV, 307
DISTR.NORM, 183
DISTR.NORM.ESTAND, 183
DISTR.NORM.ESTAND.INV, 183
DISTR.NORM.INV, 183
Distr.T.Inv, 251
FRECUENCIA, 22
MEDIA.ACOTADA, 89
MEDIA.ARMO, 49, 89
MEDIA.GEOM, 50, 89
MEDIANA, 53, 89
MODA, 56, 89
PERCENTIL, 89
Poisson, 159, 329
Promedio, 46
PROMEDIO, 89
PROMEDIOA, 89
prueba.Chi, 338
PRUEBA.CHI.INV, 317
prueba.F, 309
Funciones de Excel, 89

g

Garbage in, garbage out, 9
Generación de números aleatorios en Excel, 212
GIGO *véase Garbage in, garbage out*
Grados de libertad, 249
Gráfica de probabilidad normal, 428
Gráficas, 2
circulares, 31
de control, 32
de pastel *véase* Gráficas circulares
en Excel, 26
para series de tiempo, 478
Gráfico de probabilidad normal, 428

h

Hipótesis
alternativa, 270
estadística, 269
nula, 269
Histogramas, 26, 30
Hoja de trabajo de Excel, 22
Homoscedasticidad, 412

i

Independencia estadística, 132
Indexes, 469
Índice(s)
agregado, 450
bursátiles, 468
de la bolsa, 3
de Laspeyres, 454
de Paasche, 455
de precios agregados, 453
de Precios y Cotizaciones, 401, 469 *véase* IPC
de Precios y Cotizaciones de la Bolsa Mexicana de Valores, 48
Dow Jones, 450
en cadena, 459
estacionales, 492
ideal de Fisher, 456
Industrial Dow Jones, 48
Nacional de Precios al Consumidor, 463
Nacional de Precios al Productor, 468
ponderados, 452
simple, 450
Ingreso, 384
INPC *véase* Índice Nacional de Precios al Consumidor
Instituto Nacional de Estadística y Geografía, 451, 464
Intercepción, 384
Interpretación teórica o clásica de la probabilidad, 120
Intervalo(s)
de clase, 57
de confianza, 241
de predicción, 392
Investigación de operaciones, 158
IPC, 3, 32

j

Jerarquía y percentil, 92

l

Ley de la oferta y la demanda, 374
Límite inferior de la clase modal, 57
Linealizar, 483

m

Marco muestral, 215
Matriz para ANOVA, 352

Media, 2, 45, 61, 149, 233
en una serie de clases y frecuencias, 47
en una serie de datos y frecuencias, 46
en una serie simple, 45
Media aritmética, 44, 45, 153
fórmula, 45, 170
Media armónica, 48
Media de la distribución
de probabilidad, 149, 170
hipergeométrica, 170
muestral del estadístico, 294
multinomial, 170
Media general o global, 369
Media geométrica, 50
Media móvil, 91
Mediana, 52, 61
en una serie de clases y frecuencias, 54
en una serie de datos y frecuencias, 54
Media ponderada, 47
Media y varianza de una distribución Poisson, 170
Medidas, 2
de composición, 81
de dispersión, 67
de forma, 82
de posición, 44, 45
de tendencia central, 44 *también véase* Medidas de posición
Método(s)
clásicos, 511
de mínimos cuadrados, 377, 412
de *P*, 292
de pruebas de hipótesis, 291
del estadístico de prueba, 278, 281, 292
del intervalo, 277, 278, 292
del valor crítico, 278
del valor de la *P*, 280, 282
estadístico, 4
paramétricos, 337, 511
Mínimos cuadrados, 393
Moda, 55
en una serie de clases y frecuencias, 56
procedimiento para determinar, 57
Modelo
aditivo, 476
de análisis de varianza de un factor, 369
de regresión, 386
de regresión lineal múltiple, 411
de regresión lineal simple, 431
lineal simple, 411
multiplicativo, 476
Momentos, 82
Muestra, 4, 209
aleatoria, 211
aleatoria simple, 214
estadística representativa, 209
periódica, 213
sistemática *véase* Muestra periódica
Muestras
aleatorias, 4
apareadas, 521
de juicio, 211
independientes, 214
no independientes *véase* Muestras relacionadas
relacionadas, 214
Muestreo, 2, 209
aleatorio estratificado, 214
aleatorio simple, 214
aleatorio sistemático, 214
con reemplazo, 153, 231
para control de la calidad en producción, 513
por conglomerados, 215
sin reemplazo, 153, 232
Multicolinealidad, 415

n

New York Stock Exchange, 450
Nivel
de confianza, 241
de significación, 271, 512
Número de las muestras, 214
Números índice, 3
usos principales, 450
en cadena, 460
índice simples, 451
Números pseudoaleatorios, 211
NYSE-Euronext, 469

o

Observaciones aberrantes, 428
Organización de datos estadísticos, 5

p

Parámetro, 4, 240
Parámetros poblacionales, 210
Parámetros y estadísticos, 210
Par de hipótesis, 293
Pendiente
estimación por intervalo, 391
F de Fisher, 390
pruebas de hipótesis, 389
t de Student, 390
Percentiles, 57
Permutaciones, 118
Pesos corrientes *véase* Valores corrientes
Pesos reales *véase* Valores reales
Población, 3, 209
Polígono de frecuencias, 30
Precio productor, 463
Presentación de datos estadísticos, 5
Primer momento, 82
Principales fuentes de datos, 5
Probabilidad, 108, 124, 148
condicional, 129, 130
de ocurrencia, 112, 120

Procedimiento para estimar una media, 242
Procedimiento de tabla de clases y frecuencias, 19
Promedio, 45
Promedio aritmético *véase* Media aritmética
Promedio móvil, 51
  exponencial, 479
Promedios de cuadrados
  del error, 370
  tratamientos, 370
Promedios móviles, 51
Pronósticos
  intervalos de confianza, 432
  con promedios móviles exponenciales, 480
Proporción, 2, 81, 233
Prueba de corridas (rachas), 513
Prueba de Friedman para diseños en bloques aleatorizados, 539
Prueba de hipótesis sobre el coeficiente de correlación, 403
Prueba de Kruskal-Wallis para más de dos medias, 536, 537
Prueba de los signos, 518
Prueba del signo
  para dos muestras apareadas grandes, 522
  para dos muestras apareadas pequeñas, 521
  para una muestra grande, 521
  para una muestra pequeña, 519
Prueba de una proporción con *z*, 321
Prueba de Wald-Wolfowitz, 513
Prueba de Wilcoxon, 526
  para dos muestras apareadas grandes, 531
  para dos muestras apareadas pequeñas, 530
  para una muestra grande, 529
  para una muestra pequeña, 527
Prueba de χ^2, 512
Prueba *F*, 419
Prueba para la diferencia entre dos proporciones con *z*, 323
Prueba piloto, 215
Pruebas de dos extremos o colas, 275
Pruebas de bondad de ajuste, 512
Pruebas de hipótesis, 2, 210
  no paramétricas, 269
  p, 293
  para proporción, 285
  sobre una proporción poblacional, 285
Pruebas no parámetricas, 512
Pruebas para dos muestras, 513
Pruebas parámetricas, 511
Pruebas para una muestra, 513
Pruebas sobre una varianza, 512
Prueba U de Mann-Whitney, 532, 533, 535
Punto medio de clase, 47

r

Racha, 514
Rango, 67, 68
Recopilación de datos, 5
Región de aceptación, 275
Región de rechazo, 275
Regla de la adición de probabilidades, 125
Regla de la multiplicación de probabilidades, 134
Regla de la probabilidad condicional, 133
Regla de las 5, 329
Regla de multiplicación, 134
Regla de Sturges, 18
Regla general de la suma de probabilidades, 127
Reglas de conteo, 116
Regresión en Excel, 383
Remesas internacionales, 29
Representatividad de la muestra, 469
Residuales estandarizados, 427

s

Salarios mínimos por profesión, 12
Segundo momento, 82
Selección previa, 438
Serie(s), 12
  de clases y frecuencias, 12, 17
  de datos agrupados, 12
  de datos y frecuencias, 12, 17
  de tiempo, 29, 476
  de valores cualitativos, 20
  de valores cuantitativos, 12
  simple, 16
Sesgo, 82, 229
SPSS, 9
Statistical Package for the Social Sciences *véase* SPSS
Suavización con promedios móviles exponenciales, 479
Subconjunto, 110
Suficiencia, 242
Suma(s)
  de cuadrados, 382
  de cuadrados del error, 370
  de cuadrados de tratamientos, 369
  de cuadrados total, 369
  de las frecuencias, 17

t

Tabla(s), 2, 12
  de datos cruzados, 21
  de datos no cruzados, 21
  de contingencias, 37
  de frecuencias, 16, 19
  de probabilidades binomiales, 154
  simples, 19
  y gráficas, 11
Tamaño de muestra necesario, 245
Tasa constante, 202
T de Student, 292, 278, 296
Técnica de regresión por pasos, 437
Técnicas de análisis, 1
Tendencia secular, 476
Teorema central del límite, 227
Teorema de Bayes, 136
Teorema de Chebyshev, 75
Teoría de colas, 158
Teoría de conjuntos, 109
Teoría de la probabilidad, 2, 108

Teoría de líneas de espera, 158
Tercer momento, 83
Tipo de datos, 2
Tratamiento, 346

U

Unidad de muestreo, 215
Unidad experimental, 346

V

Valor
  calculado del estadístico de prueba, 512
  central de cada clase *véase* Punto medio de clase
  crítico de *F*, 420
Valor de *z* para un área, 186
  de la izquierda de la media hasta la derecha de la distribución normal, 187
  entre dos valores positivos de *z*, 186
  entre dos valores negativos de *z*, 187
Valor empírico de la *F*, 420
Valor empírico del estadístico de prueba, 512, 370
Valor esperado de la variable, 150
Valores corrientes, 466
Valores reales, 466
Variable, 6
  aleatoria, 147
  aleatoria Poisson, 158
  binomial, 147
  continua, 146, 147
  cualitativas, 433
  de bloque, 348, 363
  de interés, 6
  dependiente , 363
  de respuesta, 346
  de tratamiento, 363
  discontinua, 146
  discreta, 147
  discretas, 6 *véase* Variable discontinua
  ficticias, 433
  independiente, 374
  nominales, 6
  ordinales, 7
  predictora, 416
Variaciones
  cíclicas, 476, 500
  estacionales, 476
Variaciones estacionales, 492
Variaciones irregulares, 476
Varianza, 67, 149, 233
  en una serie de clases y frecuencias, 73
  en una serie de datos y frecuencias, 72
  de la distribución de probabilidad, 170
  de una población
  fórmula, 170, 231
  modificada, 231, 233